Hatred Behind the Screens
A Report on the Rise of Online Hate Speech

Professor Matthew Williams of HateLab, Cardiff University and Mishcon de Reya

A+ MISHCON ACADEMY

Hatred Behind the Screens
A Report on the Rise of Online Hate Speech

Professor Matthew Williams of
HateLab, Cardiff University
and Mishcon de Reya

Foreword

by James Libson, Executive Partner, Mishcon de Reya

With the depth of real-world divisions, and the power of the internet to amplify as well as anonymise, online hate speech is growing, with potentially profound consequences for victims and society at large. Tackling it raises not just practical issues, but fundamental questions about how much we are willing to police and curb freedom of speech, and whom we hold to account, in particular platforms - such as Twitter and Facebook - that host third-party content. The terms "hate crime" and "hate speech" are not even legally defined. Instead, what we have is a patchwork of offences, across multiple statutes, protecting certain characteristics (including race and religion) but not others, such as gender and age.

This report shines a spotlight on the nature and scale of online abuse, as well as the legal framework and ongoing initiatives to tackle it. A cornerstone of our work at Mishcon de Reya is to help our clients fight online abuse. We support the growing push in the UK and elsewhere, for much tougher internet regulation in order to address illegal content – including hate speech - and to shift responsibility onto Big Tech. We will be providing our views to the Law Commission's comprehensive review of hate crime legislation.

However the law alone cannot solve this urgent, global problem. What we need is a multi-faceted approach, including better training for police officers and prosecutors, as well as more commitment to the powerful and promising use of counter-speech.

Working with Professor Matthew Williams and his HateLab at Cardiff University, we seek not only to monitor online hate speech and its effects, but to collaborate on further research that - we hope - will form part of the solution.

"No one is born hating another person because of the colour of his skin, or his background, or his religion. People must learn to hate, and if they can learn to hate, they can be taught to love, for love comes more naturally to the human heart than its opposite."
Nelson Mandela, Long Walk to Freedom, 1994

CONTENTS

Executive Summary 8

Introduction 10

SECTION ONE
THE PROBLEM OF ONLINE HATE
Definitions of Online Hate Speech 14
Words that Wound: The Harms of Online Hate Speech 17
Patterns of Online Hate Victimisation 21
Trigger Events and Online Hate Speech 24
Bots and Fake Accounts 29
High-profile Targets of Online Hate 30

SECTION TWO
HATE SPEECH – THE LEGAL LANDSCAPE
International Instruments 34
Overview of UK Domestic Law
— Criminal Law 35
— Civil Law 38

SECTION THREE
DEVELOPING SOLUTIONS TO ONLINE HATE
Counter Speech 40
Operational Initiatives 43
Independent Internet Regulator 43
Legislative Reform 44

Conclusion 45

Note on Methodology 46

Appendix
The Existing Law Relating to Online Hate 48

Footnotes 68

Sources 72

Executive Summary

— The reporting, recording and incidence of online hate speech has increased over the past two years.

— While the number of people personally targeted remains relatively low, large numbers of people are being exposed to online hate speech, potentially causing decreased life satisfaction. In particular, an increasingly large number of UK children (aged 12-15) report that they are exposed to hateful content online.

— Online hate speech tends to spike for 24-48 hours after key national or international events such as a terror attack, and then rapidly fall, although the baseline of online hate can remain elevated for several months. Where it reaches a certain level, online hate speech can translate into offline hate crime on the streets.

— Hate crime, including hate speech, is both hard to define and hard to prosecute. A patchwork of hate crime laws has developed over the last two decades, but there is concern the laws are not as effective as they could be, and may need to be streamlined and/or extended - for example to cover gender and age-related hate crime. The Law Commission is currently reviewing hate crime legislation, and has separately completed a preliminary review of the criminal law in relation to offensive and abusive online communications, concluding there was "considerable scope for reform".

— According to a recent survey by Demos, the public appreciates the difficult trade-off between tackling hate crime and protecting freedom of speech, with 32% in favour of a safety-first approach, 23% in favour of protecting civil liberties, and 42% not favouring either option.

— Hate speech often spreads via social media and other "intermediary" platforms, currently largely shielded from legal liability as "platforms not publishers". There is however a growing push for new and greater internet regulation, recently echoed by Facebook itself. The DCMS and Home Office in their April 2019 draft White Paper on Online Harms proposed a new statutory duty of care for tech companies, to be overseen by an independent internet regulator with considerable powers, including to impose substantial fines. The European Commission is drawing up similar plans for a comprehensive "Digital Services Act" to tackle illegal online content, which may involve setting up a centralised EU regulator.

— We agree with the Government's stance in its Online Harms White Paper and would welcome a statutory duty of care, enforced by an independent regulator, to make companies take more responsibility for the safety of their users and deal with harm caused by content or activity on their services.

— In the meantime, counter-speech - any direct or general response to hate speech that seeks to undermine it - has been shown to be effective in the right circumstances and if delivered in the right way. Following terrorist events for example, tweets from verified media, the Government and police have been shown to gain significant traction, although counter-speech is less likely to be effective against fake or bot accounts, or those who hold the most extreme views. The effectiveness of counter-speech is being further tested and explored by HateLab at Cardiff University and other researchers across the globe.

Introduction

Online hate speech is now recognised as a pernicious social problem. According to Home Office data, 1,605 hate crimes were flagged as online offences between 2017 and 2018, a 40% increase on the previous year.[1] This mirrors a rise in annual prosecutions for online hate, up by 13% to 435 in the year to April 2018.[2] Both figures are likely significant underestimates, due to under-reporting by victims and substandard police recording practices.[3]

Research has also revealed patterns of hate speech, namely that hate crimes spike in the aftermath of certain 'trigger' events. In the UK, the Brexit vote in 2016, and a string of terrorist attacks the year after, prompted unprecedented increases in hate crime - both online and offline. Those intent on spreading hate have been emboldened and galvanised by international events, such as the election of Donald Trump and the rise of the far- and extreme-right in Europe and beyond. They have also been given a voice and reach by the internet which, as the All-Party Parliamentary Group on Hate Crime identified in its latest report, is a "key breeding ground" for radicalisation and grooming to extremist narratives.

The harms caused by online hate speech to victims and communities are not trivial, often matching those caused by physical crimes. Victims report feeling fear, anger, sadness, depression, and a newfound prejudice against the attacker's group, as well as physical effects including behavioural changes and isolation. Research also shows that online hate speech is often a pre-cursor to, or an extension of, offline hate crime, which can multiply and intensify the effects.[4]

The Government has responded with several legislative and policy initiatives, including the National Online Hate Crime Hub in 2017 to coordinate hate crime reporting and, more recently, its draft Online Harms White Paper, which promises a new, independent regulator and a legal duty of care for tech companies. The EU has also proposed a tougher approach in the form of a comprehensive 'Digital Services Act' to be unveiled at the end of 2020, which will force tech giants to remove illegal content or face fines.

At the same time, the debate around hate speech vs. free speech is far from settled. In a recent UK survey on opinions towards technology, 89% and 70% of respondents, respectively, identified extremist content and online abuse as key concerns.[5] But, when asked whether they favoured protection against harmful content vs. protecting free speech, 32% picked the former and 23% the latter i.e. 42% did not favour either option, and were presumably conflicted. [6]

In light of these difficult trade-offs, and as internet platforms develop more proactive, technological solutions, counter-speech – which is any direct or general response to hate speech that seeks to undermine it – has been heralded as a promising tool. Through appeals to reason and by calling out prejudice, counter-speech can help to neutralise hate speech and, crucially, make it less socially acceptable.

This report provides an overview of the latest evidence on online hate speech. Section One focuses on the many and challenging attempts to define hate speech, as well as observed patterns of hate speech and the impact on victims and communities. Section Two outlines the existing legislative checks on hate speech, both at home and abroad, including a series of case studies on how UK law has been applied. It also examines non-legal solutions, notably counter-speech. It concludes with our look to the future.

“NATIONAL GOVERNMENTS NOW RECOGNISE ONLINE HATE SPEECH AS A PERNICIOUS SOCIAL PROBLEM.”

Definitions of Online Hate Speech

The UK Crown Prosecution Service does not provide a specific definition of online hate speech, but they do state, when deciding on prosecution, regard must be given to whether the online speech was motivated by any form of discrimination or hostility against the victim's ethnic or national origin, gender, disability, age, religion or belief, sexual orientation or gender identity. The presence of any such motivation or hostility mean that it is more likely that prosecution is required. Beyond the UK, cultural and linguistic differences make a general legal definition of hate speech difficult to formulate. A focus on the expressive and emotional components of hate speech, as opposed to specific phrases and slurs, helps to deal with these differences. Legal academics have developed a set of expressive criteria, which if met, should qualify hate speech as criminal in their view.[7]

The speech in question should be criminal if it:

Deeply wounds those targeted

Causes gross offence to those that hear it, beyond those targeted

Has a degrading effect on social relationships within any one community

Provokes a violent response

Other academic work stresses the targeted nature of hate speech, where the normatively irrelevant characteristics of individuals or groups single them out. Hate speech then stigmatises victims, who are regarded as 'legitimate targets' in the eyes of the perpetrator.[8] While these are academic definitions, international organisations and governments have made efforts to develop conventions and laws that embody similar criteria.

Definitions of general hate speech have been developed by international organisations, including the Council of Europe. Recommendation (97)20 of the Council of Europe (1997) states "the term 'hate speech' shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin."

The full Twitter policy is available here.

In the UK broadly speaking, for hate speech to be considered illegal it must be grossly offensive and/or inciting others to hate in a threatening or abusive way. Speech that is threatening or harassing is criminal regardless of hateful content.

However, merely offensive, controversial or distasteful hate speech is likely to fall under Article 10 of the European Convention on Human Rights, which protects freedom of speech that includes the freedom to offend others. In such cases, the hate speech is unlikely to reach the bar for a criminal offence, meaning the police and the courts won't get involved, instead deferring to the website owner (see Section Two for a full overview of the law in the UK).

Twitter and Facebook, as well as many other social media platforms, have recently developed working definitions of online hate speech, and have integrated them into their terms of use.

Twitter came under criticism for not taking a firm stance on hate speech, which prompted its CEO to introduce a raft of new measures. Twitter Rules and Policies (2019) now state : "You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories."

Examples of what we do not tolerate includes, but is not limited to behavior that harasses individuals or groups of people with:

violent threats

wishes for the physical harm, death, or disease of individuals or groups

references to mass murder, violent events, or specific means of violence in which/with which such groups have been the primary targets or victims

behavior that incites fear about a protected group

repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that dehumanises, degrades or reinforces negative stereotypes

Twitter also forbids the use of hateful imagery, such as symbols associated with hate groups, images depicting others as less than human or non-human, and moving images (e.g. gifs, video) containing any of the above.

The full Facebook policy is available here.

Facebook Community Standards (2019) state: "We define hate speech as a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We also provide some protections for immigration status. We define "attack" as violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation."

> REGARD MUST BE GIVEN TO WHETHER THE ONLINE SPEECH WAS MOTIVATED BY ANY FORM OF DISCRIMINATION

Facebook separate levels of severity:
Tier 1 hate speech is that which:
— Is violent
— Is dehumanising and makes comparison to filth, bacteria, disease or faeces; to animals; to sub-humanity
— Mocks the concept, events or victims of hate crimes

Tier 2 hate speech is that which:
— Implies physical deficiencies related to hygiene or appearance
— Implies the inferiority of a person's or a group's mental or moral deficiency
— Expresses contempt or disgust
— Uses slurs directed at a person or group of people who share protected characteristics

Tier 3 hate speech is that which:
— Calls to exclude or segregate a person or group of people who share protected characteristics that do not fall under general criticism of immigration policies and arguments for restricting those policies
— Describes or negatively targets people with slurs, where slurs are defined as words commonly used as insulting labels for the above-listed characteristics

Shortly following the extreme right-wing terror attack in Christchurch, New Zealand, Facebook announced it would ban praise, support and representation of white nationalism and separatism on its platform.

These examples are provided here to demonstrate the recent moves made by social media companies to combat hate speech. Most other mainstream social media companies have rules similar to these that allow for non-criminal sanctions to be imposed, such as the removal of the hate speech, and the suspension and cancellation of user accounts. However, as these policies take force, some users migrate to fringe social media sites, such as Gab, Voat and 4chan, effectively displacing hate speech. Despite the introduction of 'soft' sanctions, several UK Government reports have actively criticised mainstream social media platforms for their lacklustre approach to hate speech.

Ahead of the publication of the Government White Paper on Online Harms (2019), the Digital, Culture, Media and Sport Committee's report 'Disinformation and 'fake news" (2019) concluded that Facebook and other companies continue to allow the publication of hate speech and propaganda and remain unable or unwilling to prevent these damaging communications. The second report of the House of Lords Select Committee on Communications (2019) noted that misuse of online personal data and hateful speech make the case for regulation compelling. Subsequently, the Online Harms White Paper outlined several recommendations for regulation.

Words that Wound: The Harms of Online Hate Speech

When taking into account the 'virtual' nature of online interaction, where action can seem devoid of consequence due to anonymity and sometimes vast distances between speakers, it is not surprising to encounter claims that hate speech is a trivial issue. Indeed, some of those who support right-wing perspectives have argued that online hate speech is less serious than hate crime in the physical world. However, to make such a claim is to grant the would-be online offender the right to attack individuals on the basis that their actions do not harm the victim or themselves.

Research on offline hate speech has found that victims experience trauma in a pattern that is similar to the response of victims of physical crimes. The short- and long-term effects of hate speech are similar in form to the effects of burglary, domestic violence, assault and robbery. Words and phrases that target a person's core identity can generate negative emotional, attitudinal and behavioural changes, all of which are exacerbated if the victim is already vulnerable (e.g. depression, anxiety, lack of support network) and if the context is conducive (e.g. culture of fear, repression or intimidation symptomatic of a pattern of similar behaviours).[9]

Short-term impacts lasting a few days, can include feelings of shock, anger, isolation, resentment, embarrassment and shame. Long-term impacts, lasting months or years, can include low self-esteem, the development of a defensive attitude and prejudices against the hate speaker's group, concealment of identity and heightened awareness of difference.[10] Remembering hate speech attacks has also been associated with increases in self-rated stress levels, and increased levels of the stress hormone, cortisol in LGBT victims.[11]

Direct online hate speech can be experienced by victims as an extension or a pre-cursor to hate crime in the offline world.[12]

Hate crime is not always a discrete isolated event, and for some victims it is experienced as a process, involving a range of incidents over time, from hate speech online to hate speech offline, and possibly face-to-face threats and physical violence.[13] Hate crime can therefore be considered as re-engineered to function in the online environment, utilising new technologies such as social media. For some, this is the beginning of a longer process of victimisation that may migrate offline, while for others, it is an isolated event that remains online. But regardless of the form of victimisation, the emotional consequences of online hate speech are felt in the offline world by those targeted and the communities to which they belong. For some, hate speech perpetrated online has similar, if not more pronounced effects than hate crime offline.[14]

'The UK Safer Internet Centre (2016) survey of young people aged 13-18 (N=1,512) found that those encountering online hate reported feeling anger, sadness and shock and a majority (74%) stated it made them modify their online behaviour.[15]

In a 2018 academic study, researchers found respondents most frequently exposed to online hate material in the US and Finland noted that they were less satisfied with their lives.'[16]

Reports from victims of homophobic online hate speech show evidence of shock, fear and anger; deterioration in mental and physical wellbeing; social isolation; fear for physical safety; heightened sense of threat; relationship breakdown; ripple effect on partners, families and friends; changes in behaviour in public; changes in use of social media; and periods of sick-leave.[17]

Online hate speech has the potential to inflict more harm due to several unique factors associated with Internet communication.

The perceived anonymity offered by the Internet means offenders are likely to produce more hate speech, the character of which is more serious, given increased disinhibition. The temporal and geographical reach of the Internet means hate has become a 24/7 phenomenon.

For many, especially young people, communicating with others online is now a routine part of everyday life, and simply turning off the computer or mobile phone is not an option, even if they are being targeted with hate.[18] Online hate speech then has the insidious power to enter the traditional safe haven of the home, generating a cycle of victimisation that is difficult to break. When individuals claim they have been injured by hate speech, they are ascribing power to language that equals the force

of some physical acts. Online hate speech is said to have an illocutionary force: an act of speaking or writing which has a tangible or real outcome.[19] Examples of illocutionary speech include a priest stating "I now pronounce you husband and wife", a police officer saying "You're under arrest on suspicion of grievous bodily harm", or a Judge saying "This court finds you guilty of murder". These words have significant weight, and some forms of hate speech can carry similar power with deeply serious consequences: So called illocutionary force in hate speech is accomplished in 5 ways:

1. Through social media posts that invoke rule infraction (e.g. a tweet containing a picture of a gay couple kissing could motivate hate speech that draws on laws in a country that criminalises homosexual relations).

2. Through social media posts that attempt to induce shame in the victim (e.g. the same tweet could motivate hate speech that uses the gaze of the victim's parents or grandparents as a vehicle for shame: "Imagine what your mother/grandmother would think if they saw this disgusting image!").

3. Through social media posts that attempt to induce fear in the victim (e.g. the use of threats, intimidation etc.).

4. Through social media posts that attempt to dehumanise the victim. (e.g. comparing individuals or groups to insects, vermin, or primates).

5. Through social media posts that attempt to spread misinformation related to the victim or the group they belong to (e.g. creating conspiracy theories or false information in relation to past events (such as the Holocaust) or religious holidays (such as Ramadan)).

These five forms of hateful illocutionary online speech are more likely to have the desired negative consequences noted above, if the conditions of uptake, context and power are met:

1. The uptake of the post by the victim. Only when the victim recognises they are being targeted because of their identity can they recognise the act as hate speech. There are situations where uptake fails, or the hate speech misfires. For example, the use of a slur that is not familiar to the victim due to cultural or temporal variation. In these circumstances, while the perpetrator of the hate speech may still be guilty of sending grossly offensive communications, the impact on the intended victim will be negligible, at least in the immediate term.

2. The context of the hate speech is conducive. When a victim is aware that they are interacting in a culture of fear, intimidation and repression where personal characteristics are routinely targeted and there are no laws protecting them, the pains of hate speech are amplified.

3. The power of the perpetrator outweighs that of the victim. When the hate speech offender is perceived by the victim to hold more power, whether that be offline or online status, they are more likely to feel subordinated. Due to the additional level of vulnerability brought about by this power difference, the victim is likely to feel the pains of hate speech more. This is certainly the case with humiliation, that occurs within relationships of unequal status where the humiliator dominates the victim and undermines their sense of identity.[20]

The effects of hate speech on the wider community are also documented. Those who witness hate speech or hear of it second-hand, while escaping the direct harm caused, do not escape its more general consequences. When members of minority groups hear of a spate of hate crimes online or in their village, town or city, their levels of stress and fear increase, resulting in behaviour changes, including avoiding leaving the home. Some non-victims report becoming prisoners in their own homes simply because of what they read online about hostility directed towards members of their community.[21] An abundance of hate speech can also impact on the behaviours of other would-be offenders. A cascade effect has been reported where online hate speech reaches a critical mass around a particular event. The results are not only found online. Social media posts containing anti-refugee and anti-Muslim content have been found to correlate with hate crimes on the streets in Germany, the US and the UK.[22]

The UK based study, conducted by HateLab, found a consistent positive association between Twitter hate speech targeting race and religion and racially and religiously aggravated offences on the streets in London. The study concludes online hate victimization is part of a wider process of harm that can begin on social media and then migrate to the physical world.

Patterns of Online Hate Speech Victimisation

Home Office analysis of all hate crimes reported to the police shows that 1,605 (2%) were flagged as online offences between 2017-18, an increase on the 1,148 in the previous year.

Online race hate crime makes up the majority of all online hate offences

However, transgender individuals are most likely to be victimised online:

6% of all transgender based hate crime is committed online, compared to

4% for disability and sexual orientation based hate crime,

3% of religious based hate crime and

2% of race based hate crime.

Crown Prosecutions Statistics show that in the year April 2017 to 2018 there were 435 offences related to online hate under section 127 of the Communications Act 2003 and section 1 of the Malicious Communications Act 1998, an increase on the 386 recorded in the previous year.[23]

Analysis of police crime data obtained from freedom of information requests shows that online hate crimes targeting people with disabilities increased by 304% between 2015 and 2017.

This represents the highest increase across all protected characteristics, including religion (240%), sexual orientation (208%), race (184%) and transgender identity (88%). [24]

Official figures should be treated with a degree of caution, due to significant underreporting of hate and online offences, and widespread poor practice in police recording. Her Majesty's Inspectorate of Constabulary and Fire and Rescue Service 2018 report 'Understanding the Difference: The Initial Police Response to Hate Crime' found that forces across the country have been slow to understand the changing nature of hate crime, and to take online offending more seriously. Despite the Home Office introducing a requirement for police forces to flag cyber-enabled hate crime offences, uptake of this practice has been patchy and inconsistent, resulting in unreliable data on this type of offence.[25]

Charities that support victims of hate crime also collate data on online hate speech. While these data do not reflect accurate information on actual numbers of criminal hate offences, they do allow us to get a sense of hate speech trends over time.

In the first six months of 2019, Community Security Trust, a charity that supports the Jewish community, recorded 323 online anti-Semitic UK based incidents, representing 36% of all incidents. This represents an increase of 46% on the same period the year before.[26] In the first six months of 2018, Tell MAMA, a charity providing support to victims of anti-Muslim hate crime, recorded 207 UK based incidents of Islamophobic online hate speech, representing 34% of all reports. The previous year online Islamophobic hate speech represented 30% of all recorded incidents.[27]

Stonewall, a charity that supports lesbian, gay, bisexual and transgender people in the UK, found in its 2017 survey that 10% of respondents had suffered direct online hate abuse. This represents an increase of 5% from the 2013 survey. Disaggregated figures for 2017 show online victimisation was significantly higher for respondents who identified as transgender (26%), non-binary LGBT (26%), young LGBT (23%) and black, Asian and minority ethnic LGBT (20%). Witnessing online LGBT hate material aimed at others varied by age and ethnicity. Overall, 45% of respondents had seen such material, with the number rising to 72% for young LGBT and 66% for black, Asian and minority ethnic LGBT respondents. In 2013, 28% of LGBT respondents had encountered online hate material.[28] The Galop 2016 LGBT+ Hate Crime Survey found that 30% of respondents had reported experiencing online LGBT+ hate crime.

Recent academic research that examined hate speech on social media found high rates of exposure in four countries.[29] A large representative survey (N=3,565) covering 15-30 year-olds in the US, UK, Germany and Finland found on average 43% of respondents had encountered hate material online. This rose to 53% for the US, while 39% of UK respondents reported encountering such material. Most hate material was encountered on social media, such as Facebook, Twitter and YouTube, and websites dedicated to self-harm. These conclusions are supported by anecdotal experience of our own Mishcon de Reya Cyber investigations into social media issues, where it is often difficult to avoid potentially hateful material.

The number of survey respondents being personally targeted by hate material was much lower, at around 11%. Although it must be noted that the sample did not specifically target those with protected characteristics.

Rates were highest in the

US (16%),
UK (12%)
Finland (10%)
Germany (4%)[30]

Similarly, rates of sending hate material were low in the sample. Those in the US were most likely to admit to this act (4%), followed by respondents in Finland (4%), the UK (3%) and Germany (1%). Young men living alone with a close connection to the online world were most likely to post hate material.[31]

A 2016 survey of young people aged 13-18 (N=1,512) in the UK found that 82% of respondents had encountered online hate speech, the majority of which was based on race, religion or sexual orientation. Around a quarter (24%) stated that they had personally experienced hate because of their protected characteristic (including gender).[32]

In 2019 Ofcom published the "Adults' Media Use and Attitudes Report" which found that 53% of internet users reported seeing hateful content online in 2018, an increase from 47% in 2017. Those aged 16-34 were most likely to report seeing this content (71% for 16-24s and 63% for 25-34s). Only 41% of those who witnessed online hate took action in relation to the most recent incident, and older respondents (55+) were less likely to act compared to younger respondents.

In February of 2019 Ofcom published their annual report "Children and parents: media use and attitudes." Since 2016 the report has asked 12-15 year-olds in the UK the following question: "In the past year, have you seen anything hateful on the internet that has been directed at a particular group of people, based on, for instance, their gender, religion, disability, sexuality or gender identity? Examples of these sorts of things might be nasty or hateful comments or images that have been posted on social media, comments in response to an article that you read online, or videos posted on sites like YouTube." Despite almost all respondents stating they had been told how to use the Internet safely, 45% of 12-15 year olds in 2018 reported encountering hateful content online, unchanged from the 2017 figure. However, in 2016 the figure was 34%. Of those that encountered this content, 42% took action, including reporting the post to the website and commenting on its inappropriateness.

Trigger Events and Online Hate Speech

The sustained increase in exposure to online hate content recorded in the Ofcom research in 2017 and 2018 may be a reflection of the number of 'trigger' events that occurred over the previous two years. 'Trigger' events are incidents that motivate the production and spread of hate both on and offline. Research has shown that hate crime and speech tend to increase dramatically in the aftermath of certain antecedent events, such as terror attacks and controversial political votes. Before and while the Ofcom survey was in the field (April-June 2017 and 2018) the Brexit vote and result occurred and the UK suffered three major terror attacks. All of these events culminated in an unprecedented rise in hate crime as recorded by the police.

'TRIGGER' EVENTS MOTIVATE THE PRODUCTION AND SPREAD OF HATE BOTH ON AND OFFLINE.

Following trigger events, it is often social media users who are first to publish a reaction.[33] Figure 1 shows the Twitter reaction to the murder of Lee Rigby in Woolwich, 2013. The maps of the UK and London show the location of tweets about the attack, with clusters appearing in Manchester (the family home of Rigby) the Midlands, South Wales and the West, the East, and Woolwich. The textual content of tweets is presented in the wordcloud, a representation of the most frequent words used across all tweets posted.

Figure 1: UK Twitter reaction to the Woolwich terror attack in 2013

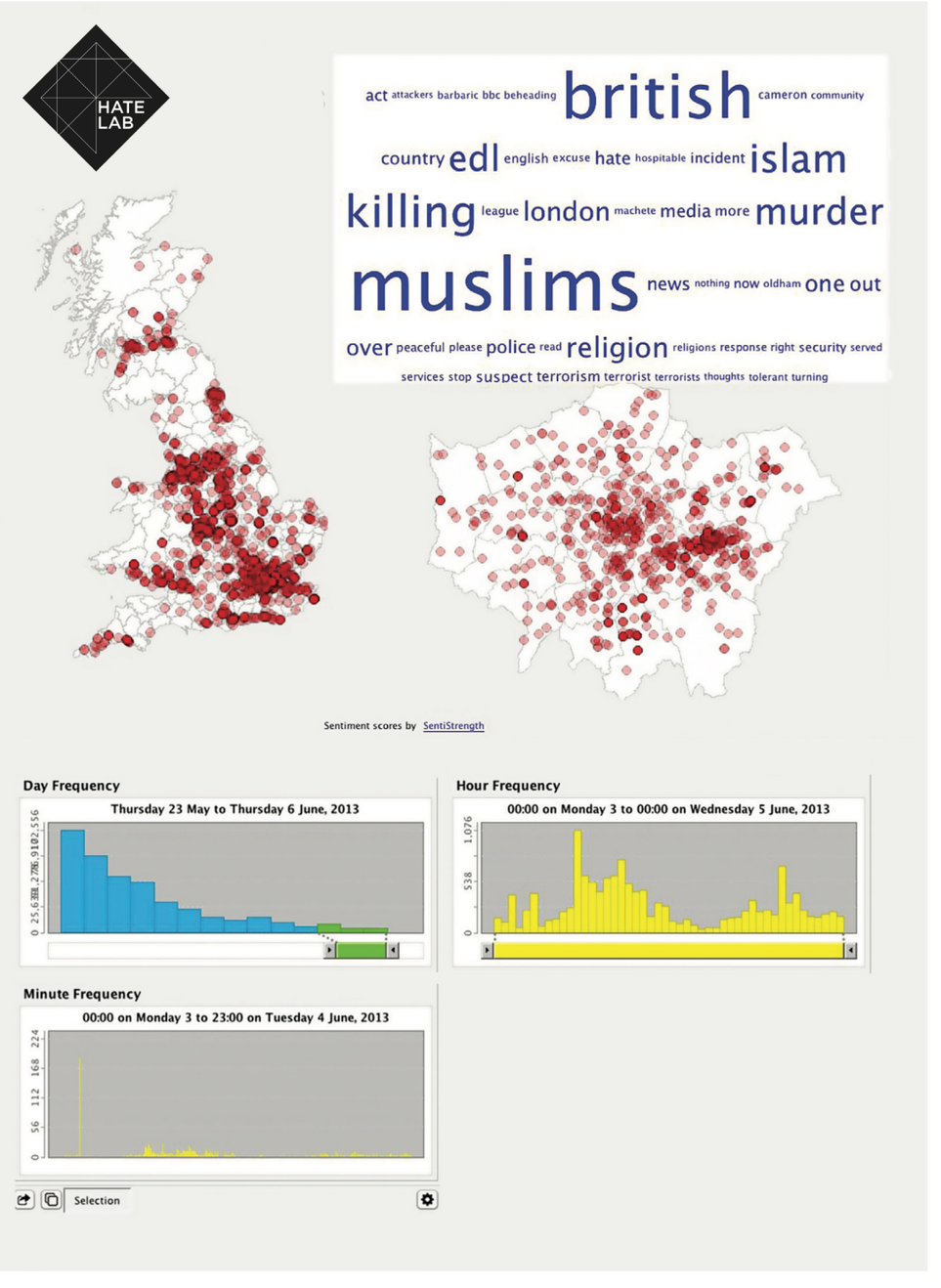
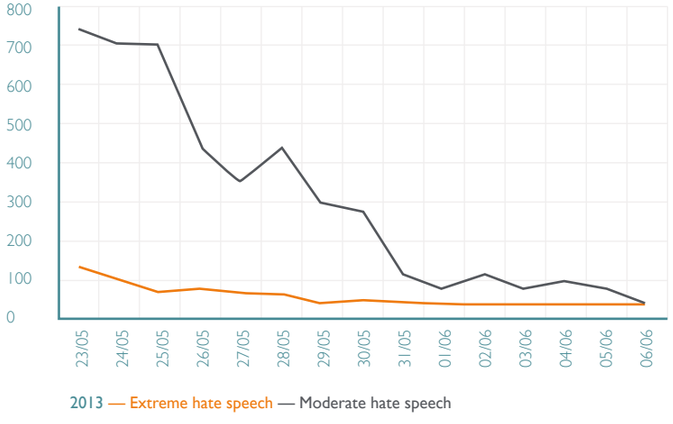


Figure 2 shows the frequency of moderate and extreme anti-Muslim online hate speech produced on Twitter in the aftermath of the attack. Moderate hate speech included posts that were likely to cause offense, such as "Told you we shouldn't have let them Muslims in. Send them home! #BurntheQuran". Extreme hate speech included similar content, but also degrading racial slurs and expletives. Both forms of online hate speech peaked the day of the attack and then rapidly declined within the first 48 hours of the aftermath. This indicates a 'half-life' of online hate.[34]

Figure 2: UK anti-Muslim hate speech on Twitter following the Woolwich terror attack in 2013



This 'half-life' is also found in relation to anti-Muslim online hate speech produced and shared following the Brexit vote in June 2016 (Figure 3) and anti-Semitic hate speech produced and spread following the breaking of the Ken Livingstone story the same year (Figure 4). In both cases hate speech spikes on or just following the date of the incident and then sharply declines. In the case of terror attacks, the rapid escalation and decay in the frequency of hate speech posts has been explained by the capacity of 'trigger' events to temporarily reduce the tendency of some users to control or regulate their implicit prejudices held towards individuals who share similar characteristics to the perpetrators. The posting of hate is further encouraged by others who post similar messages (a cascade effect) and the perception that such actions have little or no personal consequence.[35] Following a frisson of hateful sentiment in the first 24-48 hours, in the days and weeks after the 'trigger' event, users begin to regulate their implicit prejudices, and the posting of hate reduces.[36] However, what is observable in the weeks and months following these incidents is that hate speech production and propagation remains higher on average than in the proceeding period. This has led some to conclude that we are living with a new baseline of hate speech online.

Figure 3: UK anti-Muslim hate speech on Twitter around the Brexit vote

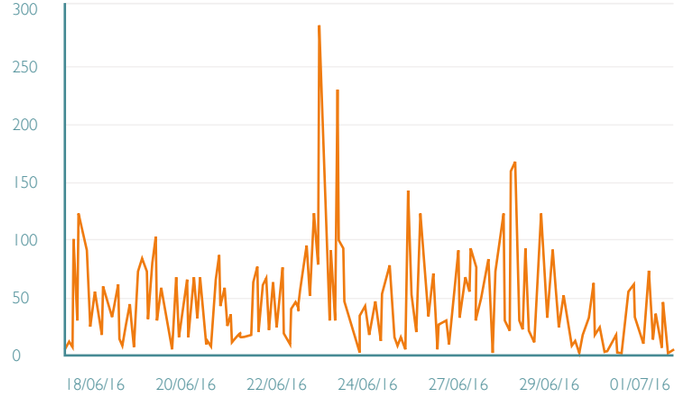


Figure 4: UK anti-Semitic hate speech on Twitter around the Ken Livingstone story breaking in 2016

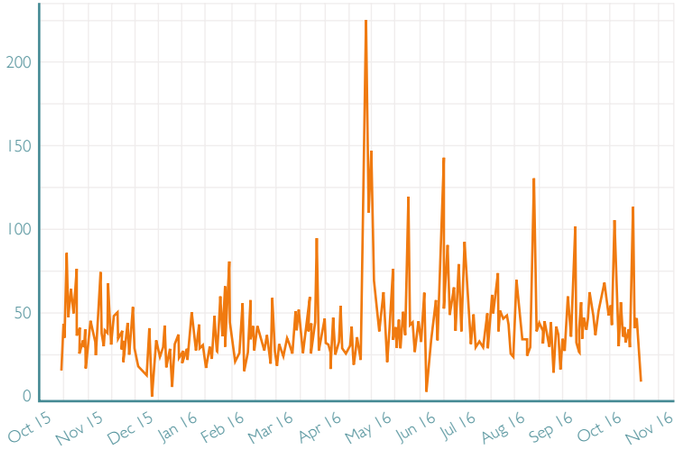
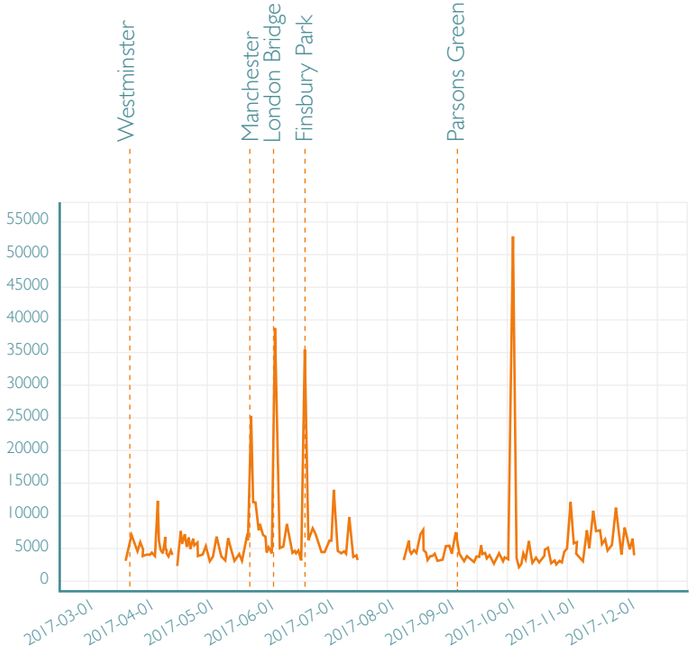


Figure 5 shows anti-Muslim hate speech posted globally on Twitter throughout 2017.[37] Discernible spikes in hate speech are evident that coincide with key events during the year, notably the UK terror attacks in Westminster, Manchester, London Bridge and Finsbury Park.[38] The large spike in October is related to the mass shooting in Las Vegas, US, where for a period of time assumptions were being made on Twitter that the attack was an extremist Islamic terrorist incident, fuelled in part by a false claim by ISIS that the shooter was acting on their behalf.

Figure 5: Global anti-Muslim hate speech on Twitter during 2017 (gaps relate to breaks in data collection)



In the events presented here, statistical models showed that those identifying with far and extreme right-wing groups were most likely to produce hateful content on Twitter. While not all hate speech reached the criminal threshold set out by the Crown Prosecution Service, some of it is was deemed sufficiently offensive to warrant requests to social media providers to delete content and user accounts for infringing platform rules.

Statistical modelling also showed that across all events, compared to all other types of online content, hate speech was least likely to be retweeted in volume and to survive for long periods of time, supporting the 'half-life' hypothesis. Where hate speech is re-tweeted following an event, there is evidence to show this activity emanates from a core group of like-minded individuals who seek out each other's messages. These Twitter users act like an 'echo chamber', where grossly offensive hateful messages reverberate around members, but rarely spread widely beyond them.[39] Hate speech produced around the Brexit vote in particular was found to be largely driven by a small number of Twitter accounts. Around 50% of anti-Muslim hate speech was produced by only 6% of users, many of whom were classified as politically anti-Islam.[40]

Tweets from verified, media, government and police accounts gained significant traction during and in the aftermath of terrorist events.[41] As these tweeters are least likely to produce and propagate hate speech, and are more likely to engage in the spread of positive messages following 'trigger' events, they proved to be effective vehicles for stemming negative content through the production of counter-speech. In particular, the dominance of traditional media outlets on Twitter, such as broadsheet and TV news, leads to the conclusion that these channels still represent a valuable pipeline for calls to reason and calm following possibly fractious events of national interest (see end of this section for a further discussion of counter-speech).

Bots and fake accounts

In October 2018 Twitter published over 10 million tweets from around 4,600 Russian and Iranian-linked fake/bot accounts. Bots are automated accounts that are programmed to retweet and post content for various reasons. Fake accounts are semi-automated, meaning they are routinely controlled by a human or group of humans, allowing for more complex interaction with other users, and for more nuanced messages in reaction to unfolding events. While not all bots and fake accounts are problematic (some retweet and post useful content) many have been created for more subversive reasons, such as influencing voter choice in the run up to elections, and spreading divisive content following national events.

Bots can sometimes be detected by their characteristics that distinguish them from human users. These characteristics include a high frequency of retweets/tweets (e.g. over 50 a day), activity at regular intervals (e.g. every five minutes), content that contains mainly retweets, the ratio of activity from a mobile device versus a desktop device, accounts that follow many users, but have a low number of followers, and partial or unpopulated user details (e.g. photo, profile description, location etc.). Fake accounts are harder to detect, given they are semi-controlled by humans, but telltale signs include accounts less than 6 months old, and profile photos that can be found elsewhere on the internet that clearly do not belong to the account.

Research at Cardiff University showed that fake Twitter accounts purportedly sponsored by overseas state actors spread fake news and promoted xenophobic messages following the 2017 terror attacks in the UK, that had the potential consequence of raising tensions between groups. Following the Manchester and London Bridge terror attacks, a fake account assumed to be linked to Russia sent a racially divisive tweet within minutes of the news breaking.[42] In the minutes following the Westminster terrorist attack, suspected fake social media accounts retweeted fake news about a woman in a headscarf apparently walking past and ignoring a victim. This was retweeted thousands of times by far-right Twitter accounts with the hashtag '#BanIslam'. All four terror attacks in the UK in 2017 saw an online response from these Russian fake accounts, with close to 500 original messages being retweeted over 150,000 times.[43] A key tactic used by these accounts was to attempt to engage celebrity and far-right accounts, thereby increasing the profile of their messages if a response was made.

The additional challenge created by these fake accounts is that they are unlikely to be susceptible to counter-speech and traditional policing responses. It therefore falls upon social media companies to detect and remove such accounts as early as possible in order to stem the production and spread of divisive and hateful content.

High-profile targets of online hate

There has been a recent trend of high-profile personalities being targeted with hateful and threatening social media posts. Examples include Gina Miller, Diane Abbott MP, Luciana Berger MP, Stella Creasy MP, Anna Soubry MP and Nicky Morgan MP, all of whom received threatening Twitter posts, some using hate speech targeting the race, religion, or gender of the victims. Many of the online attacks followed events, such as the key moments in the Brexit process. In the case of Anna Soubry and Nicky Morgan, multiple threats were received following the publication of the "The Brexit Mutineers" front page of the Daily Telegraph. Several accounts, including those thought to be fake and linked to countries outside of Europe, retweeted the headline and engaged in abusing the MPs.

Research conducted on Twitter based abuse relating to political issues (i.e. not personal characteristics) between 2015-2019 found prominent MPs, and in particular male Conservative MPs, were the most targeted, with Boris Johnson, Jeremy Hunt, David Cameron and George Osborne attracting the most abuse. The topics driving the abuse included Brexit, the NHS and terrorism. Abuse directed towards female MPs, while less frequent, was more likely to use of the "f" word and compare victims to "dirt".[44]

Nicky Morgan has cited online abuse as one of the reasons she is not standing at the General Election in December 2019, a factor also mentioned by Heidi Allen, who is also standing down as an MP. There are widespread concerns that many of the other female MPs who are not going for re-election have decided not to do so due to the abuse they have received.

Diane Abbott, the shadow Home Secretary, came under a relentless campaign of racist and sexist abuse in the weeks before the 2017 General Election. The abuse directed at her amounted to 10 times as much as was received by any other MP, according to a study by Amnesty International (2017). Luciana Berger was targeted multiple times with anti-Semitic hate speech, notably by Garron Helm in 2014, and Joshua Bonehill-Paine in 2016. Both received custodial sentences for their Twitter posts. In the former case, the incarceration of Helm resulted in a campaign of hate against the MP, with coordination linked back to the Neo-Nazi Stormfront Website.

The Committee on Standards in Public Life (2017) report 'Intimidation in Public Life' found that social media was a significant factor in amplifying intimidating behaviour towards those serving in public office in recent years, with women, Black and Minority Ethnic, Lesbian, Gay, Bisexual and Transgender and other minority parliamentary candidates being disproportionally targeted.

During the course of research we undertook at MdR Cyber we focused on the hate that MPs face day to day. We reviewed whether this hate was replicated online.

— At present much of this hate speech is directed at a small number of users. This is not to say it is not prevalent, but that for MPs it is sometimes drowned out by political influence campaigns. We worked with Astroscreen, who specialise in social media disinformation detection and who are a member of the MDR Lab incubator.

— Astroscreen and MDR Cyber took a sample of the 1.6 million tweets, and downloaded all the additional information that Twitter provides. We also collected 400 of the tweets of each of the users in the dataset (260 thousand). In the process, we ended up with a dataset of around 65 million tweets.

— We found that a selection of users were being used to amplify a variety of hashtags, often containing specific political content. If used for hate or for counterspeech these could be a powerful tools.

— We found a small selection of users generate a large amount of the tweets to MPs. Many of these are accounts set up for a single issue.

> IN THE UK, THERE IS NO SINGLE PIECE OF 'HATE CRIME' LEGISLATION

International Instruments and Initiatives

Cross-border attempts to tackle hate speech date back to at least 1965, when the United Nations adopted the International Convention on the Elimination of All Forms of Racial Discrimination, partly in response to apartheid and post-war anti-Semitism. That has since been supplemented – in 2013 – by a UN General Recommendation on Combating Racist Hate Speech, which asks participating states to criminalise behaviours including the dissemination of ideas "based on" racial or ethnic superiority or hatred; incitement offences; and participation in groups that promote and incite racial discrimination. In 2019, the UN launched a Strategy and Plan of Action on Hate Speech, with the aim of enhancing its efforts to address the root causes and drivers of hate speech, and enabling effective UN responses to the impacts on society. These instruments have moral force but limited direct impact; individual nations can still choose how and to what extent to implement broad calls for change.

The European Union has also headed several initiatives targeting hatred, and specifically online hatred, including in 2003 the Additional Protocol to the Council of Europe Convention on Cybercrime, and in 2008 the Council Framework Decision on using criminal law to combat certain forms and expressions of racism and xenophobia.

Most recently, in 2016, the EU introduced a voluntary code of conduct aimed at faster response times to illegal online hate speech, which has now been accepted by nine major web companies, including Facebook, Twitter, Instagram and YouTube. The companies have committed to reviewing the majority of content removal requests within 24 hours, then removing content as necessary. According to the latest monitoring report (February 2019), 89% of notifications by participating NGOs and public bodies are now assessed within 24 hours (up from 40% in 2016), and removals of hate speech have increased to 72% from 28% in 2016. This suggests the companies have strengthened their reporting systems, and reflects the fact they have devoted more staff and resources to content management.

The results of the code of conduct monitoring feed into the European Commission's wider work on tackling illegal content online, notably its plans to unveil a comprehensive Digital Services Act at the end of 2020, which will force tech giants to remove illegal content (including hate speech) or face fines. The Act will make social media companies subject to mandatory “notice and take down” orders - including in relation to certain types of racist and xenophobic content - and may involve setting up a centralised EU regulator. It will replace the existing 20-year-old e-Commerce Directive but apparently retain the "safe harbour" provision, which exempts online companies from direct legal responsibility for user-uploaded material.

Overview of UK Domestic Law

— Criminal Law

In the UK, there is no single piece of 'hate crime' legislation; the terms 'hate crime' and 'hate speech' are not even legally defined. In fact, the broad umbrella of 'hate crime' encompasses numerous, specific offences across multiple statutes, involving hostility aimed at one of the "protected characteristics", including race, religion and sexual orientation. The law also provides for harsher sentences where any criminal offence is aggravated by such hostility. Alongside the hostility-focused offences, there are specific legal protections against stirring up or inciting hatred. Hatred in the legal sense is more serious than hostility and must be targeted at a whole group, in such a way as to affect public order. As the CPS Guidance makes clear, the incitement offences that have been successfully prosecuted go "well beyond" voicing an opinion or causing offence.

The other key offences relating to what is commonly referred to as 'hate speech' are the communications offences, covering messages – including on social media – that are grossly offensive, indecent, obscene, menacing or false, as well as harassment, where the speech crosses the line into the unacceptable.

Finally, where the behaviour in question takes the form of speech, there are specific considerations bearing in mind the high threshold necessary to protect freedom of expression. Speech that crosses the threshold will be more than shocking or distasteful, and incitement offences in particular are highly sensitive, requiring the consent of the Attorney General in order to bring charges. Nonetheless, as the CPS also recognises, it views all hate crime seriously, given the potentially profound and lasting impact on individual victims.

Case studies

Several online hate speech cases have been brought before the courts that involve social media. The case studies below show that in some cases it is clear where a prosecution is necessary, while in others it may not be in the public interest to proceed with a prosecution.

Caroline Criado-Perez

In 2013, feminist campaigner and journalist, Caroline Criado-Perez began a petition to replace the planned image of Winston Churchill on the new £10 note, with a female figure. The campaign was a success, and the Bank of England announced than an image of Jane Austen would appear on the new note to be issued in 2017. In response to this announcement, Criado-Perez was subject to hateful comments and threats of sexual violence on social media. John Nimmo and Isabella Sorley sent death and rape threats that caused Criado-Perez to install a panic button in her home. Both pleaded guilty to sending menacing tweets, admitting they were among the users of 86 separate Twitter accounts from which Criado-Perez had received abusive messages. Before passing sentence, the judge highlighted the extreme nature of the threats and the harm caused to the victim. Isabella Sorley was jailed for 12 weeks and co-defendant John Nimmo was jailed for 8 weeks for threatening behaviour.
In 2017 Nimmo was sentenced to two years and three months in prison for sending threatening and racist emails to MP Luciana Berger. One featured a picture of a knife and the text "You are going to get it like Jo Cox". Another called Berger "Jewish scum" and was signed off "your friend the Nazi". For his offences that were racially aggravated, sentences were uplifted by 50 percent.

Fabrice Muamba

In 2012, Swansea University student, Liam Stacey targeted Bolton Wanderers football player, Fabrice Muamba, with an offensive remark on social media following him suffering a heart attack on the pitch. While Stacey's first post could be regarded as offensive, it was not directly racist: "LOL. Fuck Muamba. He's dead!!!". The offence caused other social media users to react in defence of Muamba. It was Stacey's response to these users that took on a directly racist character. Several reports from the public were made to the police and Stacey was subsequently arrested. In court, before passing sentence, the judge highlighted the racist and aggravating nature of his posts, their reach amplified by social media and the news event, and the significant public outrage that resulted. Stacey was charged under Racially Aggravated Section 4A of the Public Order Act 1986 and sentenced to serve 56 days in prison.

Dog trained to give Nazi salute

In 2018, UKIP member, Mark Meechan was found guilty under the Communications Act and fined £800 for posting grossly offensive material on YouTube after uploading a video of his girlfriend's dog trained to give the Nazi salute to the commands "sieg heil" and "gas the Jews". The video was viewed over three million times on the platform. In defending his actions, Meechan stated that he was only making a joke that was intended to annoy his girlfriend and to be consumed by those who subscribe to his YouTube channel. He apologised for causing offence. The sheriff of the court stated that the video not only showed the dog responding to a Nazi command and anti-Semitic speech, but also showed it watching a clip of a Nuremberg rally and flashing images of Hitler. By deliberately using the Holocaust as a theme for the video, Meechan was deemed to have caused gross offence that went beyond the limits of free speech. Meechan appealed the decision stating that the court's acceptance of the context of the video was at odds with what he intended. His appeal was found to lack merit and was refused.

Tom Daley and Peter Waterfield

In 2012, Port Talbot football player, Daniel Thomas sent a homophobic offensive tweet referencing Olympic divers Tom Daley and Peter Waterfield: "if there is any consolation for finishing fourth at least daley and waterfield can go and bum eachother #teamHIV". The tweet was not directly sent to Daley or Waterfield via the @ mention feature on Twitter, and hence was meant for a more general audience. Thomas was arrested and charged but not prosecuted. The Director of Public Prosecutions (DPP) decided, following consultation with Daley and Waterfield, that the tweet was not grossly offensive, was not intended to reach Daley and Waterfield, and was not part of a campaign. Therefore, the communication fell below the threshold for criminal prosecution. Thomas was also swift to remove the message and showed remorse for causing any offence. The DPP concluded that while merely offensive social media posts may not warrant criminal prosecution, they may attract alternative non-custodial sanctions, including those that can be administered by social media platforms.

Jay Davidson

In August of 2018, Jay Davidson posted images on Instagram of himself topless holding a rifle with text urging people to "stand up" next to Nazi references such as "Heil", "Aryan" and the phrase "fuck you racist Allah cunt". The post was then shared on WhatsApp. A report was made to the police and Davidson was subsequently arrested at his home. He denied being a racist and claimed there was no malicious intent behind the post. In his defence he told police he was drunk at the time and admitted to having a drinking problem. When sober the next morning he claimed feeling disgusted with himself after realising what he had done. Davidson was found guilty of stirring up religious and racial hatred and sentenced to 4 years in prison.

UK Civil Law

As in criminal law, the terms 'hate crime' and 'hate speech' do not exist in civil law. There are, however, civil remedies that are potentially available for hate speech (i.e. speech that amounts to hate crime), as well as for speech that does not reach the criminal (or at least the prosecutorial) threshold, which we might deem 'hateful speech'. These remedies might be pursued alongside or instead of criminal action, for example where an online post is defamatory, amounts to a breach of privacy or data protection laws, or constitutes harassment (a civil as well as a criminal offence).

However, much hate or hateful speech is not targeted at specific individuals and therefore not actionable; there needs to be an identifiable claimant or group of claimants, and provable loss or damage. More broadly, as with criminal hate speech but even more so in the civil sphere, arguments as to what is and is not acceptable, bearing in mind freedom of speech, are never easy.

Another key issue is that hate speech often spreads via social media and other internet “intermediaries” such as Facebook or Twitter, which are largely shielded from liability as “platforms not publishers”. They have been positively encouraged to operate a passive "notice and takedown" model, not least as the result of various exemptions – including for those “merely hosting” content - afforded by the EU E-Commerce Directive. But, this and other relevant EU legislation may cease to apply post-Brexit, and in the meantime (see below), the Government is exploring ways of shifting more legal liability onto internet platforms.

“COUNTER-SPEECH CAN HAVE A POSITIVE EFFECT BY STEMMING THE PROPAGATION OF HATE”

Counter-Speech

Counter-speech is any direct or general response to hateful or harmful speech which seeks to undermine it. Influential speakers can favourably influence discourse through counter speech by having a positive effect on the speaker, convincing him or her to stop propagating hate speech or by having an impact on the audience – either by communicating norms that make hate speech socially unacceptable or by 'inoculating' the audience against the speech so they are less easily influenced by it.

All of the studies that examined hate speech production around 'trigger' events cited earlier in this report also found the presence of counter-speech. Counter-speech is a common response to online hate speech. It can have a positive effect by stemming the propagation of hate and, when involving groups of people, reinforces norms of acceptable behaviour.[45] Combating hate speech with counter-speech has some advantages over law enforcement sanctions: i) it can be rapid, ii) it can be adaptable to the situation; and iii) it can be employed by any Internet user.

A descriptive examination of counter-speech public pages on Facebook found they were less numerous and active than their counterpart right-wing public pages. The counter-speech most likely to be produced utilised parody and satire.[46] To test if online counter conversations were useful in positively engaging with online right-wing extremists, researchers identified candidates on Facebook and engaged in outreach work. Measures of success included initial response rates, sustained engagement and indicators that the candidate was questioning their online behaviour. Results showed that 16% of candidates responded to the initial outreach contact, and that an argumentative tone produced the greatest success, followed by casual, meditative and reflexive tones. A scholarly tone yielded no response. Of these, 64% engaged in sustained conversation, and 6% had subsequently questioned their online behaviour.[47]

The Online Civil Courage Initiative (OCCI) was established in 2016. OCCI is a collaboration between the International Centre for Radicalisation and Political Violence (ICSR), the Institute for Strategic Dialogue (ISD), the Amadeu Antonio Foundation and Facebook. Its aim is to combat extremism and hate speech on the internet through the promotion of good counter-speech on social media. They draw expertise from technology, communications, marketing and academic sectors to educate and upskill civil society organisations in their counter-speech activities.

The Institute for Strategic Dialogue (ISD) report "An imprecise science: Assessing interventions for the prevention, disengagement and de-radicalisation of left and right-wing extremists" (2019) found:

- Online interventions include a broad array of approaches, including direct one-to- one outreach, one-to-many outreach, and organised campaigns.
- Unsolicited online interventions may be more effective when the practitioner is engaging with less radicalised individuals.
- Technical changes to a platform, or shifts in platform policy, have the potential to disrupt negatively the methods employed by a rangeof counterspeech practitioners.
- There is limited consensus on how to measure success in interventions due to a critical lack of systematic or independent impact evaluations.
- Intervention providers, and in particular those operating openly online, are not always fully aware of the extent of the potential risks to their own safety and wellbeing, or lack the resources or support required to mitigate these effectively.

Researchers at Cardiff University's HateLab are currently testing the effectiveness of different types of counter-speech sent to those posting hate speech on Twitter.[48] Four forms of counter speech are being considered:

Attribution of Prejudice
e.g "Shame on #EDL racists for taking advantage of this situation"

Claims making and appeals to reason
e.g. "This has nothing to do with Islam, not all Muslims are terrorists!"

Request for information and evidence
e.g. "How does this have anything to do with the colour of someone's skin??"

Insults
e.g. "There are some cowardly racists out there!"

Initial results show that counter-speech is effective in stemming the length of hateful social media conversations when multiple unique counter-speech contributors engage with the hate speech producer. However, not all counter speech is productive, and evidence shows that individuals that publicly use insults against hate speech producers often inflame the situation, resulting in the production of further hate speech. When engaging in counter-speech, or advising others on its use, the following principles should be followed to reduce the likelihood of the further production of hate speech:

Avoid using insulting or hateful speech

Make logical and consistent arguments

Request evidence if false or suspect claims are made

State that you will make a report to the police or third party if the hate speech continues and/or gets worse (e.g. becomes grossly offensive or includes threats)

Encourage others to also engage in counter-speech

If the account is likely fake or a bot, contact the social media company and ask it to be removed

While likely to be effective in some instances, general counter-speech is unlikely to stem the production of hate in social media users that are associated with the extreme right (so-called incorrigible hate offenders). Those most susceptible to the stemming effects of counter-speech are those who use hate speech only occasionally (for example, around 'trigger' events), and those that are not on a pathway to radicalisation. Counter-speech is unlikely to be effective on some bots or fake accounts, given their control is either fully or partially automated by computer code and their designed purpose is to spread hate.

Operational Initiatives

Several initiatives related to online hate speech have already been announced by the Government and are operational. The National Online Hate Crime Hub (operating through the True Vision website: http://www.report-it.org.uk/) was launched by the Home Secretary in 2017. The Hub acts a single point of contact through which all reports of online hate crime will be directed, acting in a similar way to Action Fraud, within the City of London Police. Trained officers provide guidance and specialist knowledge to police services and conduct preliminary investigations online. Police service jurisdictional issues are dealt with and, where a perpetrator can be identified, the relevant service is required to act on the investigation. The service improves victim experience of the criminal justice process that leads to more successful case building for prosecution. In addition to the Hub, the new UK Council for Internet Safety (UKCIS) has become responsible for online harms experienced by children, radicalisation and extremism, violence against women and girls, serious violence and hate crime and hate speech.

What is on the horizon?

A new, independent internet regulator

In April 2019, the Department for Digital, Culture, Media and Sport (DCMS) and Home Office published their joint White Paper on Online Harm, setting out an "ambitious vision for online safety" that goes "far beyond self-regulation". They proposed a new statutory duty of care for tech companies – defined as "companies that allow users to share or discover user-generated content or interact with each other online" - to be overseen by an independent regulator. The regulator will have a "suite of powers" to take action against companies that have breached their duty of care, including the power to issue "substantial" fines, to disrupt the business activities of non-compliant companies, and to impose liability on individual members of senior management. Companies will be required to fulfil their duty of care with reference to new codes of practice. Where these relate to illegal harms such as incitement of violence, the regulator will be expected to work with law enforcement to ensure the codes adequately keep pace with the threat.

Legislative reform

The Law Commission is currently reviewing hate crime legislation in England and Wales to test its adequacy and make recommendations for reform. It will consider whether hate crime laws should be extended to cover hostility based on age and gender - or other characteristics - and whether doing so would devalue the concept of hate crime. It will also consider the impact on reform of human rights obligations, notably rights to freedom of expression and against discrimination, likewise the implications for other areas of law, such as the Equality Act 2010. After an extensive public consultation, to begin in 2020, the Final Report is due in early 2021.

Separately, the Law Commission has produced an initial Scoping Report on the criminal law relating to offensive and abusive online communications, with a view to conducting a second-stage review. It noted that gender-based online hate crime, namely misogynistic abuse, is a particularly prevalent and damaging concern, and questioned whether the particular nature of hate speech is adequately captured in the current criminal law. It added that, where the majority of online hate speech is, in practice, prosecuted within the broader category of “grossly offensive” or “menacing” communications, the law should perhaps more explicitly address hateful communications, and label and criminalise them as such.

Conclusion

The growth of online hate speech is in many ways an example of the persistent online harms – including fake news, data capture and voter manipulation – that have prompted recent moves to regulate internet platforms as the key vehicles for those harms. Platforms are uniquely placed to remove unlawful content before it spreads, or even stop it being posted in the first place. Clearly, the major players have been slow to take responsibility for stemming hostility on their platforms, and must devote more resources to tackling the problem.

Over and above these concerns, hate speech has its own complexities, not least the problem of definition. Currently, hate crime in the UK encompasses numerous offences involving "hostility" aimed at one of five "protected characteristics". The Law Commission is considering extending that range to include age and gender, possibly others, but recognises this could undermine the very concept of hate crime, as well as impacting the criminal justice system as a whole. Last year, the head of Scotland Yard, Cressida Dick, backed a call by the chairwoman of the National Police Chiefs Council to focus on "core policing" instead of misogyny.

Meanwhile French MPs, who in July passed a law that will oblige social media networks to remove offending hate speech within 24 hours (modelled on German legislation introduced last year), faced a similar challenge to agree what constitutes "obviously hateful" material, eventually excluding references to anti-Zionism and hate against the state of Israel. Twitter was also forced earlier this year to narrow its broad prohibition on "dehumanising" speech against "identifiable groups" to just religious groups. Barring the most serious incitement offences, working out when offensive language crosses the line from merely offensive to grossly offensive, or otherwise criminal, will never be easy.

This is not to excuse social media companies; regulation aside, they can and should do more. We agree with the Government's independent Commission on Countering Extremism that platforms need to be more consistent in how they apply their own terms and conditions, including as between far right groups and other extremists; they should permanently ban those who are persistently hateful or abusive; and they should be more open with the data they hold, to help identify new and emerging trends.

Against the backdrop of this ongoing struggle to decide if and how to punish hate speech, and how to assign liability, counter-speech has emerged as a powerful tool that harnesses, rather than threatens, freedom of speech. Although it is unlikely to change the behaviour of true extremists, and has no impact on fake accounts or 'bots', counter-speech is a means of helping people to call out suspect claims, and to rally support from other moderate voices. That is especially vital when – in the rush to embrace to "de-platforming" and "safe spaces" - we seem to be losing the art of reasoned debate. To quote Federal appeals court judge Barrington D. Parker after he ruled that President Trump was not entitled to block fellow Americans from following him on Twitter, "… the best response to disfavoured speech on matters of public concern is more speech, not less."

We are at the start of a long road to understanding the full nature and scale of online hate speech, and ultimately finding the most effective solutions. Where hate speech is both a cause and symptom of societal division, part of the answer is surely addressing wider social issues. You cannot properly tackle hate speech without tackling hate; if you do, you run a greater risk of cementing hostility and driving it underground. Another part is education, particularly of the young, including in how best to deploy counter-speech. Finally, we welcome the review of existing legislative powers, which should then inform any careful changes to the legal and regulatory framework. As society grapples with online harms, we aim to be part of refining its response.

Methodological note:
Not all academic work on online hate speech has been included in this report. Decisions on what studies were included were based upon the established academic standards of rigour and significance. Several published studies were excluded due to issues with data quality (e.g. samples were not representative of the population under study, meaning estimates of victimisation prevalence were unreliable or the data presented were not collected in line with accepted ethical standards). Other studies were excluded as their findings simply corroborated existing work referenced in the report.

APPENDIX
THE EXISTING LAW RELATING TO ONLINE HATE

INTERNATIONAL INSTRUMENTS

United Nations International Convention

United Nations International Convention on the Elimination of All Forms of Racial Discrimination (1968) and the 2013 a General Recommendation on Combating Racist Hate Speech (Recommendation No. 35, CERD 2013):

Recognises racist hate speech:
used by individuals or groups
in spoken, published, and electronic forms (e.g. social media).
using hate symbols

Acknowledges perpetrators may use indirect forms of racist language to disguise hate speech, especially when attempting to appear moderate to attract support for their viewpoint

The Recommendation states the following behaviours should be criminalised:

The spread of hateful ideas

Inciting others to hate

Threatening others in the context of hate, or inciting others to do the same

Offensive hateful speech that is motivated by inciting others to hate

Membership of hate-related groups that incite hatred

Article 7 addresses the causes of hate and suggests many ways of stamping out hate speech at its core in schools, workplaces, law enforcement, the judiciary, and the public sector.

Techniques for addressing online hate speech:
Legislation that governs the operation of social media and Internet providers within State jurisdictions, drawing on international standards

Accountable social media and Internet providers that impress upon their users their responsibility for disseminating ideas and opinions

Adoption of professional ethics by social media and Internet providers that incorporate respect for the principles of the Convention and other fundamental human rights standards

Self-regulation and compliance with codes of ethics by social media and Internet providers, as underlined in the Durban Declaration, including:

Using Internet technology in the fight against racism, xenophobia and intolerance by promoting equality and non-discrimination Rapid and coordinated international response to online hate speech.

'Upon inspection of the UK's 21st to 23rd periodic reports submitted to the UN Committee on the Elimination of Racial Discrimination, a recommendation was drawn related to the increase in online hate speech following the referendum on the UK's future within the European Union. England, Northern Ireland, Scotland and Wales, as well as the overseas territories and Crown dependencies should adopt "comprehensive measures to combat racist hate speech and xenophobic political discourse, including on the Internet, particularly with regard to the application of appropriate sanctions, and ensure that public officials not only refrain from such speech but also formally reject hate speech and condemn the hateful ideas expressed so as to promote a culture of tolerance and respect" (CERD 2016: 4). The UK Government are yet to respond this recommendation.'

Council Framework Decision 2008/913/JHA

2008 Framework Decision on combating certain forms and expressions of racism and xenophobia harmonises legislation throughout the European Union. In relation to online hate speech, the Framework Decision states member States shall criminalise:

Speech that incites racist or xenophobic hatred or violence made via information systems

Speech that condones, denies or grossly trivialises crimes against humanity, war crimes and genocide that also incites hatred or violence made via information systems

The public distribution of pictures or other material via information systems in the commission of either of the above acts

'National laws in many States remain inadequate, particularly in relation to speech that condones, denies or grossly trivialises crimes against humanity, war crimes and genocide that also incites hatred or violence. England and Wales has no law criminalising this conduct, but prosecution may be brought if such speech is sent via email, social media or other online means and is considered grossly offensive or is of an indecent, obscene or menacing character.'

Additional Protocol to the Council of Europe Convention on Cybercrime

2003 Council of Europe Convention on Cybercrime Additional Protocol Concerning the Criminalisation of Acts of a Racist and Xenophobic Nature Committed Through Computer Systems sets out a comprehensive series of online activities that relate to online hate speech:

Distributing or making available racist and xenophobic material to the public through a computer system

Threats or insults that express hate towards race, colour, descent or national or ethnic origin, as well as religion sent via a computer system

Speech that condones, grossly minimises, or justifies genocide or crimes against humanity sent via a computer system

To date 44 States have signed the Protocol, with ratifications from 31. The UK neither signed nor ratified the Protocol.

European Commission Initiatives

In 2016 the European Commission, Facebook, Microsoft, Twitter and YouTube signed up to the code of conduct on countering illegal hate speech online[49], with Instagram, Google+, Snapchat, Webedia and Dailymotion joining later. IT company signatories of the code agree to:

Have rules or community guidelines banning the promotion of incitement to violence and hateful conduct and to introduce mechanisms, including dedicated teams, for the removal of illegal content within 24 hours of notification using relevant rules, guidelines and laws

Raise awareness with their users and staff about banned forms of hate speech

Streamline the notification process with EU members states and police services, to ensure timely and effective notifications of hate speech to be made

Facilitate the up-take of notification systems via partnerships with civil society organisations, who can become “trusted reporters” that can notify social media companies of cases of hate speech that contravene relevant laws, rules and community guidelines

Intensify relationships with civil society organisations to share best practice on countering hate narratives to assist in campaigns

Promote independent counter-hate-narratives, new ideas and initiatives and support educational programs that encourage critical thinking.

Intensify best practice sharing between social media companies

Evaluations on the progress made by Internet companies against the code of conduct have been conducted, and the most recent in 2018 shows 89% of companies reviewed the majority of notifications sent to them within 24 hours, and 72% of these posts were removed.[50] In 2017, 81% of companies reviewed the majority of notifications sent to them within 24 hours, and 70% of these posts were removed. In 2016, when monitoring first began, 40% of companies reviewed the majority of notifications sent to them within 24 hours, and 28% of these posts were removed. In each round of monitoring all companies but Twitter have increased removal rates. In the 2017 round Twitter removed 46% of flagged posts compared to 44% in the 2018 round. For all companies, the UK removal rate in the 2018 round was 54% (a decrease on the 66% in the 2017 round), compared to a rate of 88% in Germany in 2018 (a decrease on the 100% in the 2017 round). In the 2018 round Xenophobia (including anti-migrant hatred) was the most commonly reported grounds of hate speech (17.0%) followed by sexual orientation (15.6%) and anti-Muslim hatred (13.0%).[51]

In 2018, the European Commission published its recommendation on measures to effectively tackle illegal online content, including hate speech. The document outlines a set of common tools to ensure illegal content is swiftly detected, removed and prevented:

Detection and notification: points of contact should be established to facilitate detection and rapid take-down of content and investment should be made in algorithmic detection

Effective removal: illegal content should be removed as fast as possible, with fixed timeframes established for deeply harmful content. Transparency reports should be published detailing the number and types of notices received. Safeguards should also be introduced to reduce the risk of over-removal

Prevention of re-appearance: measures should be taken to dissuade users from repeatedly uploading illegal content. Algorithms should be developed to prevent re-appearance.

UK DOMESTIC LAW

Criminal Law

Hate Crime

The term ‘hate crime’ can be used to describe a range of criminal behaviour carried out by one or more perpetrators, such as verbal abuse, intimidation, threats, harassment, assault and bullying, as well as damage to property. Although many people and organisations use the term 'hate crime', the legal definitions focus on the word 'hostility', not 'hate'.

The law protects people against discrimination, prejudice and hostility directed towards disability, ethnicity, gender identity, nationality, race, religion or sexual orientation, where it is linked to criminal conduct. These are aspects of a person’s identity described in the law on equality as ‘protected characteristics’.

The following definition has been agreed between the Police and the Crown Prosecution Service (CPS) for identifying cases involving hostility based on protected characteristics:

‘Any criminal offence which is perceived by the victim or any other person, to be motivated by hostility or prejudice, based on a person’s disability or perceived disability; race or perceived race; or religion or perceived religion; or sexual orientation or perceived sexual orientation or a person who is transgender or perceived to be transgender.‘

This definition is used for flagging cases; for a conviction to receive enhanced sentencing in court, there needs to be sufficient evidence to prove the hostility element.

There is no legal definition of hostility, so the CPS uses the everyday understanding of the word, which includes ill-will, spite, contempt, prejudice, unfriendliness, antagonism, resentment and dislike.

In short, the law treats hostility as an aggravating feature when the hostility is:

linked to a criminal offence
in some way about one of the protected characteristics

The legal framework for hate crime prosecutions is provided by the Crime and Disorder Act 1998 (CDA 1998) and the Criminal Justice Act 2003 (CJA 2003). These two Acts operate differently but deal with the issue of hostility in a similar way. Both Acts also provide for longer and more severe sentences when hostility is present.

Specific offences

The CDA 1998 contains a number of specific offences of racially and religiously aggravated crime based on the offences of wounding, assault, damage, stalking, harassment and threatening or abusive behaviour. To prove that such offences are racially or religiously aggravated, the prosecution has to prove the ‘basic’ offence, and the racial or religious aggravation, as defined in section 28 CDA 1998.

Aggravated offences

The CJA 2003 gives the court power to enhance the sentence of any offence that is racially or religiously aggravated (section 145) or aggravated by reason of disability, sexual orientation or gender identity (section 146).

The relevant provisions within the CJA 2003 and CDA 1998 use the same terminology in setting out aggravation:

at the time of committing the offence or immediately before or after doing so, the offender demonstrated towards the victim hostility based on the victim's membership (or presumed membership) of a (specified group(s)); or

the offence was motivated (wholly or partly) by hostility towards members of a (protected characteristic) based on their membership (or presumed membership) of that (specified groups(s)).

Case law guidance

Evidence of words (spoken or written) or actions that show hostility towards the victim will be required. "Demonstrations" of hostility often involve swear words, for example: "black bastard" (R v Woods [2002] EWHC 85) or "African bitch" (R v White [2001] EWCA Crim 216). In RG & LT v DPP [2004] EWHC 183 May LJ said: "It may be possible to demonstrate racial hostility by, for instance, holding up a banner with racially offensive language on it."

In R v Rogers (2007) 2 W.L.R. 280, the defendant was involved in an altercation with three Spanish women during the course of which he called them "bloody foreigners" and told them to "go back to your own country". The House of Lords, in upholding the defendant's conviction, held that the definition of a racial group clearly went beyond groups defined by their colour, race, or ethnic origin. It encompassed both nationality (including citizenship) and national origins. The House of Lords added that the fact that the offender's hostility was based on other factors in addition to racist hostility or xenophobia was irrelevant.

The demonstration of hostility need not be based on any malevolence towards the group in question. Disposition at the time is irrelevant: see DPP v Green [2004] EWHC 1225 (Admin.) and R v Woods, in which it was irrelevant that the offender, who used racially abusive language to a doorman after being refused admission, might well have abused anyone standing in the victim's place by reference to any obvious physical characteristic.
The motivation based on hostility need not be the sole or main motivation for the offence; it may also be motivated by other reasons. In DPP v McFarlane [2002] EWHC 485 (Admin), the defendant shouted threatening and racist abuse at the victim after finding the victim parked in a disabled bay in which the defendant was entitled to park. It was immaterial that the defendant may have had an additional reason for uttering the racial words in question.

The victim's reaction to the hostility is not relevant. See R v Woods, in which the victim was called a "black bastard" but said in evidence that he was "not bothered" by such comments. The Administrative Court found that the use of racist abuse during the commission of the basic offence made out the test for racial aggravation.

How does 'hate speech' relate to hate crime?
Evidence of hostility might include words or actions at the time of the offence, or just before or after it happened.

Words might be abusive towards the personal characteristic or presumed personal characteristic, and action or behaviour might, for example, target something specific to the personal characteristic or presumed personal characteristic, such as a hijab, a yarmulke or a mobility aid. In some instances, the speech itself may amount to both a crime and the hostility element.

Stirring up hatred
Stirring up racial and religious hatred, and hatred based on sexual orientation, are offences under the Public Order Act 1986 (POA 1986), but the legal elements are different:

Stirring up racial hatred is committed when someone says or does something (including posting material online, displaying a poster, performing a play or broadcasting on the media) which is threatening, abusive or insulting, and the person either intends to stir up racial hatred or makes it likely that racial hatred will be stirred up.

Stirring up religious hatred or hatred on the grounds of sexual orientation is committed if a person uses threatening words or behaviour or displays any threatening written material (including posting material online, displaying a poster, performing a play or broadcasting on the media), and intends to stir up religious hatred or hatred on the grounds of sexual orientation.

Note that the threshold is higher for the latter set of offences: "threatening words or behaviour" versus "threatening, abusive or insulting" for racial hatred, and a likelihood of stirring up hatred is not enough. Further, only the latter set of offences contain an express freedom of expression clause to balance the right to free speech (see the section on Article 10, below) with the duty of the state to protect the rights of others and to act proportionately in the interests of public safety to prevent disorder and crime (although Article 10 is relevant to all offences). By way of example:

section 29J POA 1986 provides that, as to stirring up religious hatred, nothing in the Act "... prohibits or restricts discussion, criticism or expressions of antipathy, dislike, ridicule, insult, or abuse of particular religions, or the beliefs or practices of its adherents."

section 29JA provides that, as to stirring up hatred on the grounds of sexual orientation "for the avoidance of doubt, the discussion or criticism of sexual conduct or the urging of persons to refrain from or modify such conduct or practices shall not be taken of itself to be threatening".

Stirring up hatred means more than just causing hatred, and is not the same as stirring up tension. It must amount to hatred against a whole group – rather than hostility to just one person – and manifest itself in such a way that public order might be affected. The offences that have been successfully prosecuted go well beyond the voicing of an opinion or the causing of offence.

As CPS Guidance makes clear, "When considering whether or not to prosecute stirring up offences, there is a need to bear in mind that people have a right to freedom of speech. It is essential that in a free, democratic and tolerant society, people are able to exchange views, even when these may cause offence." The issues involved in such cases are highly sensitive and charges for stirring up hatred require the consent of the Attorney General in addition to the consent of the CPS.

Communications offences
The Communications Act 2003 (CA 2003) applies only to messages sent via a public electronic communications network, and section 127 covers the sending of improper messages. Section 127(1)(a) relates to a message that is grossly offensive or of an indecent, obscene or menacing character. Section 127(2) targets false messages and persistent misuse intended to cause annoyance, inconvenience or needless anxiety; and includes someone who persistently makes silent phone calls.

If a message sent is grossly offensive, indecent, obscene, menacing or false, it is irrelevant whether it was received; the offence is one of sending.

The Malicious Communications Act 1988 (MA 1998), section 1, deals with the sending to another of any article which is indecent or grossly offensive, or which conveys a threat, or which is false, provided there is an intent to cause distress or anxiety to the recipient. The offence covers letters, writing of all descriptions, electronic communications, photographs and other images in a material form, tape recordings, films and video recordings.

The offence is one of sending, delivering or transmitting, so there is no requirement for the article to reach the intended recipient.

Social media
Where social media is used to facilitate a substantive offence, such as a threat to kill or blackmail, prosecutors can proceed under that substantive offence; otherwise, one of the communications offences may be appropriate, and the CPS has produced specific guidance on cases involving social media communications.

Article 10
Article 10 of the European Convention on Human Rights ("ECHR") provides that:

Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This Article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises.

The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary.

However, Article 17 (often referred to as the "abuse clause") provides that:

“Nothing in this Convention may be interpreted as implying for any State, group or person any right to engage in any activity or perform any act aimed at the destruction of any of the rights and freedoms set forth herein or at their limitation to a greater extent than is provided for in the Convention.”

In other words, the right to freedom of expression under Article 10 does not extend so far as to protect an attack – for example by way of hate speech – on others' rights under the ECHR.

According to CPS Guidance, prosecutors should only proceed with cases under s.1 MA 1998 or s.127 CA 2003 if they are satisfied that there is sufficient evidence that the communication in question is more than (i.e. crosses the high threshold necessary to protect freedom of expression, even unwelcome freedom of expression):

Offensive, shocking or disturbing; or

Satirical, iconoclastic or rude comment; or

The expression of unpopular or unfashionable opinion about serious or trivial matters, or banter or humour, even if distasteful to some or painful to those subjected to it; or

An uninhibited and ill thought out contribution to a casual conversation where participants expect a certain amount of repartee or “give and take”;

This is with reference to “contemporary standards… the standards of an open and just multi-racial society", assessing whether the particular message in its particular context is "beyond the pale of what is tolerable in society" adopting the observations, as guidance illuminating these terms, in DPP v Collins [2006] UKHL 40 and Smith v ADVFN [2008] 1797 (QB).

Case law guidance
The test for "grossly offensive" was stated by the House of Lords in DPP v Collins to be whether the message would cause gross offence to those to whom it relates (in that case ethnic minorities), who need not be the recipients. The case also confirms that it is justifiable under ECHR Art 10(2) to prosecute somebody who

has used the public telecommunications system to leave racist messages.

The European Commission has held that extreme racist speech is outside the protection of Article 10 because of its potential to undermine public order and the rights of the targeted minority: Kuhnen v Germany 56 RR 205.

The ECtHR has confirmed that Holocaust denial or revision is removed from the protection of Article 10 by Article 17: see Lehideux and Isorni v France [2000] 30 EHRR 665; and M'Bala M'Bala v France (application no. 25239/13), which ruled that a blatant display of hatred and anti-semitism disguised as an artistic production (comic performance), even if satirical or provocative, was not protected by Article 10.

Prosecutors must be satisfied that a prosecution is required in the public interest and that, where Article 10 is engaged, on the facts and merits of the particular case it has convincingly been established that a prosecution is necessary and proportionate. Particular care must be taken where a prosecution is contemplated for the way in which a person has expressed themselves on social media.

Prosecutors therefore should, where relevant, have particular regard to:

The likelihood of re-offending. The spectrum ranges from a suspect making a one- off remark to a suspect engaged in a sustained campaign against a victim;

The suspect's age or maturity. This may be highly relevant where a young or immature person has not fully appreciated what they wrote;

The circumstances of and the harm caused to the victim, including whether they were serving the public, whether this was part of a coordinated attack ("virtual mobbing"), whether they were targeted because they reported a separate criminal offence, whether they were contacted by a person convicted of a crime against them, their friends or family;

Whether the suspect has expressed genuine remorse;

Whether swift and effective action has been taken by the suspect and/or others for example, service providers, to remove the communication in question or otherwise block access to it;

Whether the communication was or was not intended for a wide audience, or whether that was an obvious consequence of sending the communication; particularly where the intended audience did not include the victim or target of the communication in question.

Whether the offence constitutes a hate crime (which may mean Article 10 is not engaged, but may also be a factor tending in favour of a prosecution in the public interest).

Harassment
The Protection from Harassment Act 1997 ("PHA") creates both criminal (section 2) and civil (section 3, see below) liability for pursuing a course of conduct that amounts to harassment, and which the person pursuing the course of conduct knows or ought to know amounts to harassment (section 1(1)).

A course of conduct will amount to harassment if a reasonable person in possession of the same information as the person pursuing that course of conduct would think it amounted to or involved harassment (section 1(2)). References to harassing a person include alarming the person or causing the person distress (section 7(2)). A course of conduct requires conduct on at least two occasions (section 7(3)).

"Conduct" includes speech (section 7(4)), and the PHA can cover harassment by publication, including online.

It is a defence (section 3(1)) to show that the course of conduct:

- was pursued for the purpose of preventing or detecting crime;
- was pursued under any enactment or rule of law or to comply with any condition or requirement imposed by any person under any enactment; or
- was, in the particular circumstances, reasonable.

The Court of Appeal in R v N(Z) [2016] EWCA Crim 92 approved the following summary (per Simon J in Dowson (and others) v Chief Constable of Northumbria Police [2010] EWHC 2612 (QB)) of what needs to be proved as a matter of law to found a claim in harassment:

There must be conduct which occurs on at least two occasions,

which is targeted at the claimant*,

which is calculated in an objective sense to cause alarm or distress, and

which is objectively judged to be oppressive and unacceptable.

What is oppressive and unacceptable may depend on the social or working context in which the conduct occurs.

A line is to be drawn between conduct which is unattractive and unreasonable, and conduct which has been described in various ways: 'torment' of the victim, 'of an order which would sustain criminal liability'.

*Although, per Levi and another v Bates and others [2015] EWCA Civ 206, a harassment claim can also be brought by other persons who were foreseeably and directly harmed by the course of conduct.

The gravity of misconduct must be sufficient to sustain criminal liability. Where the definitions of civil and criminal harassment are the same, no conduct can be civilly actionable unless it is also sufficiently serious to warrant a criminal sanction (Majrowski v Guy's and St Thomas NHS Trust [2006] UKHL 34). Lord Nicholls explained:

"…courts will have in mind that irritations, annoyances, even a measure of upset, arise at times in everybody's day-to-day dealings with other people. Courts are well able to recognise the boundary between conduct which is unattractive, even unreasonable, and conduct which is oppressive and unacceptable. To cross the boundary from the regrettable to the unacceptable the gravity of the misconduct must be of an order which would sustain criminal liability under section 2."

Civil Law

There are a number of civil law offences that may apply to speech that meets the criminal standard, in addition to or instead of pursuing criminal action. Likewise, speech that falls short of that standard may still be actionable, or at least there may be grounds for formal complaint, including to the primary or secondary publisher (including platform) or regulator. However, as with criminal hate speech but even more so in the civil sphere, arguments as to what is and is not acceptable, bearing in mind freedom of speech, are never easy.

As in criminal law, the terms 'hate crime' and 'hate speech' do not exist in civil law. There are however civil remedies that are potentially available for hate speech (i.e. speech that amounts to hate crime), as well as for speech that does not reach the criminal (or at least the prosecutorial) threshold, which we might deem 'hateful speech'.

In a civil (tort) claim, there must be a claimant or group of claimants who have suffered loss or damage. A claim cannot be brought, for example, by (or on behalf of) all black people or any one black person in relation to a defamatory comment made about all black people; there is no identifiable claimant.

Much hate or hateful speech is not targeted at specific individuals and is therefore not actionable under civil law. But where, for example, abusive words about Chinese people are also targeted at a specific Chinese individual, the following civil offences may apply.

Harassment

The offence

As noted above, harassment is both a criminal and civil offence. The same defences apply to each. To recap, the PHA creates liability for pursuing a course of conduct that amounts to harassment, and which the person pursuing the course of conduct knows or ought to know amounts to harassment. In a civil claim, damages may cover any anxiety and/or financial loss caused by the harassment (section 3(2)).

The formulation of the civil offence is slightly different in that an "actual or apprehended" breach can be the subject of civil proceedings. It is therefore possible to bring a civil harassment claim where harassment is expected, but has not yet taken place. In these circumstances, a civil injunction may be ordered to protect a person who "is or may be a victim" of harassment, and breach of such injunction – without reasonable excuse – is a criminal offence.

Harassment and hate speech
To reiterate, conduct can include speech, including online speech, and two or more instances of speech can constitute a course of conduct amounting to harassment. Where the speech is "oppressive and unacceptable", and has caused another person "alarm" or "distress", that person may have a claim in harassment, and may be awarded damages (including for anxiety) and an interim or final injunction.

Where the gravity of misconduct must be sufficient to sustain criminal liability, in theory, a claimant can choose to make a civil and/or a criminal harassment complaint (possibly alongside, or instead of, action in respect of another criminal offence).

Defamation
The Offence
A statement is defamatory if it:

- identifies the claimant (this is possible as part of a class, if the words complained of would reasonably be understood to refer to each member but, the wider the class, the less likely that any one individual could establish identification);

- is published to a third party i.e. to at least one person other than the claimant;

- lowers the claimant in the estimation of right-thinking members of society/the public; and

- has caused or is likely to cause serious harm (in the case of a company, serious financial loss) to the claimant's reputation.

Any or all of the author, editor and publisher can be held liable for any defamatory/otherwise unlawful content. To date, most internet intermediaries (e.g. Facebook, Blogger), have argued that they are platforms not publishers, and therefore not liable, at least not until after they are made aware of problematic material posted by a user, at which point they may be liable for the continued publication of the material.

Alongside defences relating to the means of publication and type of publisher, such as website operators, there are defences relating to the content complained of, including:

Truth (section 2 Defamation Act 2013): if a statement is substantially true, this is a complete defence to a defamation claim.

Honest opinion (section 3 Defamation Act 2013): This applies where (a) the statement complained of is a statement of opinion (as opposed to fact); (b) the statement indicates the basis of the opinion; and (c) an honest person could have held the opinion based on any true or privileged fact that existed at the time the statement was published. The defence is defeated if the claimant can show the defendant did not hold the opinion.

Public interest (section 4 Defamation Act 2013): This statutory defence is based on the old common law 'responsible journalism' defence. It protects statements on a matter of public interest, in circumstances where the publisher reasonably believed (at the time) that publishing the statement was in the public interest, taking into account all the circumstances.

The court will only in rare cases grant an interim defamation injunction (i.e. to restrain a defendant from publication pending trial), given the public interest in freedom of speech and the principle that damages are a sufficient remedy.

Defamation and Hate Speech
The law of defamation relates to reputation, and to the particular reputation of a person or company, whereas hate speech is typically broader in its focus as well as abusive, as opposed to reputationally damaging. A statement such as "Kill all Christians", for example, does not affect the reputation of Christians, nor, as discussed above, would there be an identifiable claimant where the group is so large.

One can however imagine scenarios in which hate or hateful speech would overlap with defamatory speech and/or constitute harassment e.g. "X is one of those nasty Muslim paedophiles [false, defamatory], he lives at Y address and we should set fire to his house and make his life a living hell [hate speech and/or harassment where there is a course of conduct]".

Misuse of Private Information

The Offence
There is a specific tort of misuse of private information. The test is whether, in relation to any piece of information, the subject has a reasonable expectation of privacy. Factors to consider include (but are not limited to):

the nature of the information

the level of intrusion

the extent of previous disclosures (i.e. the extent to which the information is already in the public domain)

the number of people who know the relevant information

the steps taken to preserve privacy

The right to privacy protects not only the secrecy of private information (confidentiality), but also against intrusion.

Disclosure of private information may be justified in the public interest. When determining whether there has been a misuse of private information, the Court will balance the claimant's right to a private life under Article 8 ECHR and the defendant's right to freedom of expression under Article 10 ECHR.

The remedies available are damages and injunctive relief, including an interim injunction where private information is about to be disclosed.

Misuse of private information and hate speech
It is not often that hate or hateful speech will include information that is private. In the example given above however, disclosing a person's address to the general public would likely constitute a breach of privacy. Disclosing a person's sexuality, in the context of homophobic abuse, would also likely give rise to a privacy claim.

Unlike in a defamation claim, the truth or falsity of a statement is not relevant. So, a claimant could seek an injunction to prevent the publication of (true) details of his affair, on the basis that those details are private and that publication would be intrusive.

Regulatory and Quasi-legal Offences
Hate or hateful speech may also fall foul of specific regulations and standards, and give rise to a formal complaint on that basis.

Ofcom, for example, regulates TV and radio broadcasters licensed in the UK. It is able to impose fines, issue warnings, suspend and/or remove licences. The potentially relevant sections of its Broadcasting Code include:

Rule 2.1: "Generally accepted standards must be applied to the contents of television and radio services and BBC ODPS [BBC On Demand Programme Services] so as to provide adequate protection for members of the public from the inclusion in such services of harmful and/or offensive material."

Rule 2.3: "In applying generally accepted standards broadcasters must ensure that material which may cause offence is justified by the context [see specific definition of "context"]. Such material may include, but is not limited to, offensive language, violence, sex, sexual violence, humiliation, distress, violation of human dignity, discriminatory treatment or language (for example on the grounds of age, disability, gender reassignment, pregnancy and maternity, race, religion or belief, sex and sexual orientation, and marriage and civil partnership). Appropriate information should also be broadcast where it would assist in avoiding or minimising offence."

Rule 3.1: "Material likely to encourage or incite the commission of crime or to lead to disorder must not be included in television or radio services or BBC ODPS."

Under Rule 3.1, “material” may include but is not limited to:
• content which directly or indirectly amounts to a call to criminal action or disorder;
• material promoting or encouraging engagement in terrorism or other forms of criminal activity or disorder; and/or
• hate speech which is likely to encourage criminal activity or lead to disorder.

Hate speech is defined as "all forms of expression which spread, incite, promote or justify hatred based on intolerance on the grounds of disability, ethnicity, gender, gender reassignment, nationality, race, religion, or sexual orientation".

Rule 3.2: "Material which contains hate speech must not be included in television and radio programmes or BBC ODPS except where it is justified by the context."

Broadcasters’ attention is drawn to sections 22 and 29F of the Public Order Act 1986, which sets out criminal offences arising from the broadcast of material stirring up hatred relating to race, religion, or sexual orientation.

As for the internet, there is no single or specific regulator. In the UK, we have a patchwork of statutory and non-governmental organisations that regulate behaviour associated with the internet, such as the Information Commissioner’s Office and the Advertising Standards Authority, as well as regulators of industry-specific content, such as the Financial Conduct Authority. Currently, internet intermediaries are likely to escape primary liability for unlawful content, not least as a result of various exemptions from liability for damages - such as the "hosting" exemption - afforded by the EU E-Commerce Directive 2000. But, post-Brexit, the

Directive may cease to apply, and the Government is exploring ways, notably via an independent internet regulator, of shifting more legal liability onto internet platforms. The EU also plans, by way of a wide-ranging Digital Services Act, to introduce a centralised regulator, with the power to fine tech companies that fail to remove illegal online content, including hate speech.

At the same time, the scope of existing EU regulation is widening. A revised Audiovisual Media Services Directive (AVMSD) has been formally adopted by the European Parliament and Council and EU Member States have 21 months to transpose the new Directive into national legislation. Crucially, the revised directive has been extended to cover video-sharing platforms, such as Netflix, YouTube and Facebook, as well as standalone parts of newspapers' websites that feature audiovisual programmes or user-generated videos. The new Directive specifically states (Article 6) that Member States "shall ensure by appropriate means that audiovisual media services provided by media service providers under their jurisdiction do not contain any incitement to hatred based on race, sex, religion or nationality".

Again, Brexit creates some uncertainty. The Government has said recently that, if there is no deal, the AVMSD will no longer apply to the UK, in particular the country of origin principle, according to which AVMS providers are only subject to the jurisdiction of one EU country (the country of origin).

Finally, victims of hate or hateful speech should consider platforms' own rules and standards (see Section One).

FOOTNOTES

1. The data for 2018-19 were not made available in the Home Office Hate Crime, England and Wales, 2017/18, report. The following reason was provided: "In April 2015, it became mandatory for all forces to return quarterly information on the number of crimes flagged as being committed online (in full or in part). There are some large variations in the proportion of offences flagged by each force depending on crime type and there is anecdotal evidence to suggest that the flag is currently underused. Due to the ongoing development of the statistics and concerns around the quality of the data they have been badged as Experimental Statistics. The 'Hate Crime, England and Wales, 2017/18' bulletin included some exploratory analysis of the number of hate crimes that had been flagged as online. The analysis showed that only two per cent of hate crime offences had an online flag, which was likely to have been an underestimate and therefore any conclusions drawn from the data were done so with caution. Due to the uncertainty around the quality of the data, the analysis has not been repeated in this bulletin."

2. The CPS Hate Crime Report 2018-19 did not contain statistics on prosecutions for online hate crimes.

3. HMICFRS (2018) found that despite the Home Office introducing a requirement for police forces to flag cyber-enabled hate crime offences, uptake on this practice has been patchy and inconsistent, resulting in unreliable data on prevalence.

4. Awan and Zempi 2017, Williams et al. 2019

5. Demos 2017a

6. Preferences were made on a five-point scale. Percentages represent the top two (protection from harmful content) and bottom two (protection of freedom of speech) selections combined. The neutral selection represented 42% of respondents.

7. Greenawalt 1989

8. Parekh 2012

9. Leets 2002

10. Ibid.

11. Crowley 2013

12. Awan and Zempi 2017

13. Williams, et al. 2019

14. Galop 2017

15. UK Safer Internet Centre (2016) Creating a better internet for all: Young people's experiences of online empowerment + online hate, London: UK Safer Internet Centre.

16. Keipi, T., Räsänen, P., Oksanen, A., Näsi, M., & Hawdon, J. (2018). Exposure to online hate material and subjective wellbeing: A comparative study of American and Finnish youth. Online Information Review, 42(1), 2–15.

17. Galop 2017

18. Williams 2006, Brown 2018

19. Butler 1997

20. Silver et al. 1986, Klein 1991

21. Awan and Zempi 2017

22. Muller and Schwarz 2017, 2018, Williams et al. 2019

23. CPS 2018

24. Oliveres 2018

25. HMICFRS 2018

26. CST 2019.

27. Tell MAMA 2018

28. Bachmann and Gooch 2017

29. Hawdon et al. 2017.

30. Kaakinen 2018a

31. Kaakinen et al. 2018b

32. UK Safer Internet Centre 2016

33. Williams and Burnap 2016

34. Williams & Burnap 2016

35. In the psychology literature, this phenomenon is commonly referred to as deindividuation.

36. Williams & Burnap 2016

37. Represents both moderate and extreme anti-Muslim hate speech combined, and original tweets and retweets combined. Thanks go to my research associate, Sefa Ozalp for this graph.

38. Similar patterns were found around the Brussels, Orlando, Nice, Normandy, Berlin and Quebec attacks (Demos 2017b)

39. Williams and Burnap 2016, 2018

40. Demos 2017b

41. Williams and Burnap 2016, 2018

42. Crest 2017

43. Crest 2017

44. Greenwood, M.A. et al. (2019) Online Abuse of UK MPs from 2015 to 2019. Working paper. University of Sheffield.

45. This does not apply to counter-narratives used in an attempt to de-radicalise those engaging in extremist Islamic discourse or terrorism (see Carthy et al. 2017)

46. Demos 2015

47. ISD 2018

48. Williams & Burnap 2017

49. https://edri.org/files/privatisedenf/euhatespeechcodeofconduct_20160531.pdf

50. The 2018 figures include Google and Instagram that joined in 2018.

51. For a full breakdown see: https://ec.europa.eu/info/sites/info/files/code_of_conduct_factsheet_7_web.pdf

SOURCES

APPG (2019) How do we Build Community Cohesion when Hate Crime is on the Rise?, London: All-Party Parliamentary Group on Hate Crime.

Awan, I. and Zempi, I. (2017) 'I Will Blow Your Face Off'—Virtual and Physical World Anti-Muslim Hate Crime', British Journal of Criminology, 57(2): 362–380.

Bachmann, C. & Gooch, B (2017) LGBT in Britain, Hate Crime and Discrimination, London: Stonewall.

Brown, A. (2018) 'What is so special about online (as compared to offline) hate speech?', Ethnicities, 18(3): 297-326.

Burnap, P. and Williams, M. L. (2015) 'Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making', Policy & Internet, 4: 223–42.

Butler, J. (1997) Excitable Speech: A Politics of the Performative, London: Routledge.

Carthy, S., O'Hara, D. and Sarma, K (2017) Counter Narratives for the Prevention of Violent Radicalisation: A Systematic Review of Targeted Interventions, London: The Campbell Collaboration.

CERD (2013) General recommendation No. 35: Combating racist hate speech, UN Committee on the Elimination of Racial Discrimination, Geneva: United Nations.

CERD (2016) Concluding observations on the twenty-first to twenty-third periodic reports of United Kingdom, UN Committee on the Elimination of Racial Discrimination, Geneva: United Nations.

Crest (2017) Russian influence and interference measures following the 2017 UK terrorist attacks, Lancaster: Centre for Research and Evidence on Security Threats.

Crowley, J. P. (2013) 'Expressive Writing to Cope with Hate Speech: Assessing Psychobiological Stress Recovery and Forgiveness Promotion for Lesbian, Gay, Bisexual, or Queer Victims of Hate Speech', Human Communication Research, 40(2): 238-261.

Committee on Standards in Public Life (2017) Intimidation in Public Life, London: Her Majesty's Stationary Office.

CPS (2018) Hate Crime Report 2017-18, London: Crown Prosecutions Service.

CST (2018) Antisemitic Incidents January-June 2018, London: Community Security Trust.

Digital, Culture, Media and Sport Committee (2019).

Disinformation and 'fake news': Final Report, London: DCMSC

Demos (2015) Counter-speech: examining content that challenges extremism online, London: Demos

Demos (2017a) Public Views on Technology Futures, London: Demos

Demos (2017b) Anti-Islamic Content on Twitter, London: Demos

Eichhorn, K. (2001) 'Re-in/citing Linguistic Injuries: Speech Acts, Cyberhate, and the Spatial and Temporal Character of Networked Environments', Computers and Composition, 18: 293–304.

Galop (2017) Online Hate Speech Report 2017, Galop: London.

Greenawalt, K. (1989), 'Speech Crime & the Uses of Language', Oxford: Oxford University Press.

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A Cross-National Consideration. Deviant Behavior, 38(3), 254–266. DOI: 10.1080/01639625.2016.1196985. Journal IF 2016 [JCR]: 1.052.

Home Office (2019) Online Harms White Paper, London: Home Office.

Home Office (2017) Hate Crime, England and Wales, 2017/18, London: Home Office.

HMICFRS (2018) Understanding the Difference: The Initial Police Response to Hate Crime, London: Her Majesty's Inspectorate of Constabulary and Fire and Rescue Service.

ISD (2018) Counter Conversation: A model for direct engagement with individuals showing signs of radicalisation, London: Institute for Strategic Dialogue.

Kaakinen, M., Keipi, T., Oksanen, A., & Räsänen, P. (2018a) How Social Capital Associates With Online Hate Victimization. Policy & Internet.

Kaakinen, M., Räsänen, P., Näsi, M., Minkkinen, J., Keipi, T., & Oksanen, A. (2018b). Social capital and online hate production: A four country survey. Crime, Law & Social Change, 69(1), 25–39.

Keipi, T., Kaakinen, J., Oksanen, A. and Räsänen, P. (2017) 'Social Tie Strength and Online Victimization: An Analysis of Young People Aged 15–30 Years in Four Nations', Social Media + Society, 1–12.

Keipi, T., Räsänen, P., Oksanen, A., Näsi, M., & Hawdon, J. (2018). Exposure to online hate material and subjective wellbeing: A comparative study of American and Finnish youth. Online Information Review, 42(1), 2–15. DOI: 10.1108/OIR-05-2016-0133. Journal IF 2016 [JCR]: 1.534.

Leets, L. (2001) 'Responses to Internet Hate Sites: Is Speech Too Free in Cyberspace?', Communication Law and Policy, 6: 287–317.

Leets, L. (2002) 'Experiencing Hate Speech: Perceptions and Responses to Anti-Semitism and Anti-gay Speech', Journal of Social Issues, 58(2): 341-361.

Levin, B. (2002) 'Cyberhate: A Legal and Historical Analysis of Extremists' Use of Computer Networks in America', American Behavioral Scientist, 45: 958–88.

Muller, K. and Schwarz, C. (2017) 'Fanning the Flames of Hate: Social Media and Hate Crime', Working Paper, University of Warwick.

Muller, K. and Schwarz, C. (2018) 'Making America Hate Again? Twitter and Hate Crime under Trump', Working Paper, University of Warwick.

Ofcom (2019), Adults' Media Use and Attitudes Report, Ofcom https://www.ofcom.org.uk/__data/assets/pdf_file/0011/113222/Adults-Media-Use-and-Attitudes-Report- 2018.pdf

Ofcom (2016) Children and parents: media use and attitudes, Ofcom https://www.ofcom.org.uk/__data/assets/pdf_file/0034/93976/Children-Parents-Media-Use-Attitudes-Report-2016.pdf

Oliveres, V (2018) Tackling Online Disability Hate, Available here: https://social.shorthand.com/VictriaVic/ngRFHeHj3x/tackling-online-disability-hate.

Parekh, B. (2012). 'Is there a case for banning hate speech?' In M. Herz & P. Molnar (Eds.), The content and context of hate speech: Rethinking regulation and responses (pp. 37–56). Cambridge: Cambridge University Press.

Perry, B. and Olsson, P. (2009) 'Cyberhate: The Globalisation of Hate', Information & Communications Technology Law, 18: 185–99.

Tell MAMA (2018) Gendered Anti-Muslim Hatred and Isamophobia, London: Tell MAMA.

UK Safer Internet Centre (2016) Creating a better internet for all: Young people's experiences of online empowerment + online hate, London: UK Safer Internet Centre.

Williams, M (2006) 'Virtually Criminal: Crime, Deviance and Regulation Online', London: Routledge.

Williams, M. L. and Burnap, P. (2016) 'Cyberhate on Social Media in the Aftermath of Woolwich: A Case Study in Computational Criminology and Big Data', British Journal of Criminology, 56: 211–38.

Williams, M. L. and Burnap, P. (2017) Centre for Cyberhate Research & Policy: Real-Time Scalable Methods & Infrastructure for Modelling the Spread of Cyberhate on Social Media, ESRC Research Grant.

Williams, M. L. and Burnap, P. (2018) Antisemitic Content on Twitter, London: Community Security Trust.

Williams, M. L.et al. 2019. Hate in the machine: Anti-black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. British Journal of Criminology (10.1093/bjc/azz049).